Date of publication DEC-31, 2022, date of current version OCT-18, 2022.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.21.4.2785

Diabetes Prediction Using Binary Grey Wolf Optimization and Decision Tree

LAYLA A. AL.HAK
College of Science, University of Diyala, Iraq

Corresponding author: Layla A. Al.Hak (e-mail: laylaaeabdalhaq@uodiyala.edu.iq)

ABSTRACT Type 2 diabetes is a well-known lifelong condition disease that reduces the human body's ability to produce insulin. This causes high blood sugar levels, which leads to different complications, including stroke, eye, cardiovascular, kidney, and nerve damage. Although diabetes has attained the attention of huge research, the classification performance of such medical problems utilizing techniques of machine learning is quite low, primarily due to the class imbalance and the presence of missing values in data. In this work, we proposed a model using binary Grey wolf optimization (GWO) and a Decision tree. The proposed model is composed of preprocessing, feature selection, and classification. In preprocessing, that is responsible for minority class oversampling and handling missing values. In the second step, binary GWO are used to select the most significant features. In the third step, the proposed model is trained using the Decision tree algorithm. The model achieved an accuracy of 83.11% when it was applied on the Pima Indian`s dataset.

KEYWORDS Diabetes; Decision tree; Grey Wolf Optimization.

I. INTRODUCTION

Chronic illnesses are often considered as one of the leading causes of mortality and disability globally. Diabetes is then identified as one of the main issues of the twenty-first century [1]. More than 415 million people between the ages of 20 and 79, according to the International Diabetes Federation, have diabetes, and by 2040, that number is expected to reach 642 million. It is important to remember that this estimate does not include anyone beyond 80 [2].

Diabetes indicates that the glucose level is higher than the normal level on a regular basis. The diabetes appears when the body is unable to produce enough insulin to maintain glucose levels stable. Diabetes can be controlled with oral prescriptions, the use of insulin infusions, a restricted dietary regimen, and constant physical activity, but no complete cure exists yet [3, 4].

As is well known, until a few years ago, specialists clinically diagnosed diabetes merely on experience and with the support of laboratory-tested data or clinical information [5]. These laboratories reviewed test results that varied depending on meals, workout, illness, anxiety, minimal temperature difference, various equipment employed, and sample handling methods. As a result, the diagnosis of this type of illness is not only time-consuming but is also entirely dependent on the perception and availability of the doctors who must deal with the patient's imprecise and ambiguous clinical information [6, 7].

So, in order to improve precision with the current laboratory information while also reducing the time spent, a smart classification approach is required, which will need some input data (patients data, i.e., pima Indian Diabetes (PID)). Based on the dataset analysis, an appropriate model is built using the class description utilizing the dataset features. Depending on the same concept, that can be easily expanded, and a classification system may be constructed.

The paper is structured as follows: Section 2 addresses Related work, Methodology is presented in Section 3, proposed method is shown in Section 4, and Experimental results are given in Section 5. Conclusion is detailed in Section 6.

II. RELATED WORK

There exists a lot of research related to diabetes classification using Pima Indian's dataset. Some of them are briefly discussed below.

In [6], it was suggested a framework that utilized Principal Dimensionality Reduction and PSO for feature selection and different classification algorithm such as logistic Regression and K-Nearest Neighbor (KNN). The model was tested on two dataset PIDD and d Localized Diabetes Dataset.

In [8], Multilayer perceptron (MLP), Logistic Regression, and KNN for the classification of diabetes were utilized. The authors compared and evaluated all of the used classification algorithms where knn attained the best result in compared to MLP and logistic regression.

In work [9], a decision tree, genetic algorithm(GA), and synthetic minority oversampling technique (PMSGD) for diabetes classification on Pima Indians Diabetes Database (PID) dataset were used. The model consists of four layers. Missing values and oversampling are handled in preprocessing layer. In the second layer, important attributes are selected using a genetic algorithm and correlation.

In [10], three machine learning algorithms for diabetes mellitus classification using a support vector machine, and naive Bayes (NB) were applied. The experiment was conducted on the PID dataset. Comparatively to SVM and decision tree, Naive Bayes attained the highest accuracy.

In [11], RST-BatMiner was presented, which relies on removing the redundant attributes from the PIDD dataset, and fuzzy rules generated using the bat optimization algorithm(BOA). Further, the suggested fitness function is reduced using BOA, and features were selected using rough set theory (RST).

In [12], a method for diabetes early diagnosis was proposed, several machine learning methods were utilized for classification: Naive Bayes (NB), logistic regression (LR), and random forest (RF). The model experimented on PIDD dataset and highest accuracy is achieved by RF in comparison to the other DM method utilized.

In work [13], a method was presented to predict the types of diabetes mellitus, related future risks, and the type of medication given based on the patient's risk level. Hadoop Map Reduce environment's machine learning approach was applied to find patterns and find missing values in the Pima Indian Diabetes dataset.

III. DIABETES

Diabetes mellitus sometimes referred to as diabetes, is a metabolic condition that raises blood sugar levels. Insulin is a hormone that transports sugar from the blood into humans' cells, where it can be stored or utilized as fuel. When patient have diabetes, body either produces insufficient insulin or struggles to utilize the insulin it produces [14].

A. SYPTOMS OF DIABETES TYPE 2

The type2 symptoms can be included [15]:

- blurry vision
- hunger increase
- tiredness
- thirst increase
- urination increase
- sores that heal slowly

B. RISK FACTORS OF TYPE2 DIABETES

The risk factors are as follows[16]:

- **The weight.** If patient gains more fatty tissue, patient's cells become more insulin resistant.
- **Inactivity.** Patient's risk increases as patient becomes less active.
- **Family History.** The risk of diabetes type2 increases if a sibling or parent has type 2 diabetes.
- **Age.** As patient get older, patient's risk increases.
- **Gestational Diabetes** Patient is more likely to develop prediabetes and type 2 diabetes if patient had gestational diabetes while pregnant.
- **Polycystic Ovary Syndrome**. The risk of developing diabetes is higher for women who have polycystic ovary syndrome, a prevalent disorder marked by obesity, excessive hair growth, and irregular menstrual cycles.
- **High Blood Pressure.** An increased risk of type 2 diabetes is associated with blood pressure readings exceeding 140/90 mm Hg.
- **Abnormal Triglyceride and Cholesterol Levels.** The risk of type 2 diabetes is increased if patient's high-Density Lipoprotein, or "good" Cholesterol, levels are low.

IV. METHODOLOGY

A. DATASET

Dataset was acquired from the UCI Machine learning Repository (UCI repository of bioinformatics databases https://www.ics.uci.edu/mlearn/MLRepository.html).Table 1 illustrates the description on PIDD dataset, which is noted below.

Table 1. Pima Indian Dataset Description

No. of attributes	name of attribute	No. of instance	No. of classes
8	Triceps Skin fold thickness	768	2
	plasma glucose concentration		
	diabetes pedigree function		
	No. of Times Pregnant		
	Age		
	body mass index		
	2h serum insulin		
	diastolic blood pressure		

B. GREY WOLF OPTIMIZATION ALGORITHM

Mirjalili et al. [17] introduced the GWO, a novel metaheuristic algorithm that simulates the hunting mechanism and social hierarchy of grey wolves in nature and relies on three key steps: surrounding prey, hunting, and attacking the prey. To mathematically describe the wolf leadership hierarchy, let us name the best answer as alpha, and the third and best solutions as delta and beta, respectively [18]. The remaining possible solutions are all considered to be omega. Fig. 1 depicts the grey wolf's rigid social.

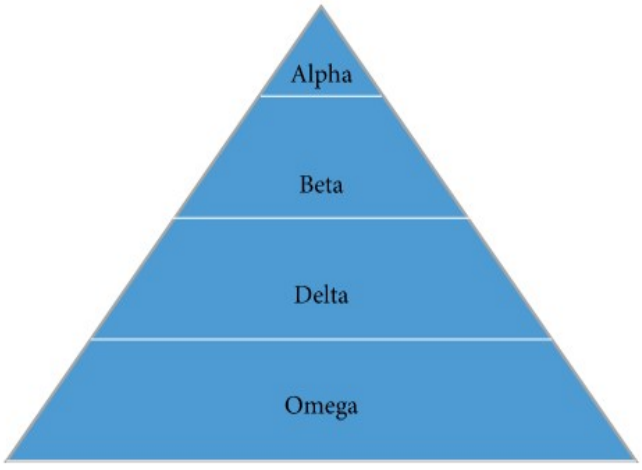

Figure 1. Grey wolves Hierarchy.

During the hunt, grey wolves encircle victims. The following equations are used to mathematically mimics the encircling behavior of grey wolves:

$$\vec{D} = \left|\vec{C}.\vec{X}_{prey}(t) - \vec{X}_{wolf}(t)\right|, \quad \vec{X}_{wolf}(t+1) = \vec{X}_{prey}(t) - \vec{A}.\vec{D}' \tag{1}$$

where $\vec{A}$ *and* $\vec{C}$ are coefficients of vectors and computed as follows:

$$\vec{A} = 2\vec{a}.\overrightarrow{r_1} - \vec{a}, \tag{2}$$

$$\vec{C} = 2\overrightarrow{r_2}, \tag{3}$$

where $\overrightarrow{r_2}$ *and* $\overrightarrow{r_1}$ are in the interval [0,1].

Alpha usually guides the hunt. Delta and beta may also join in hunting sometimes. To mathematically simulate the grey wolf hunting behavior, the first three optimum solutions (beta, alpha, and delta) acquired so far are kept, and the (omega) remaining search agents are required to update their locations according to equations (4)– (10) [19]:

$$\vec{D}_{alpha} = \left|\vec{C}_1.\vec{X}_{alpha} - \vec{X}\right|, \tag{4}$$

$$\vec{D}_{beta} = \left|\vec{C}_1.\vec{X}_{beta} - \vec{X}\right|, \tag{5}$$

$$\vec{D}_{delta} = \left|\vec{C}_1.\vec{X}_{delta} - \vec{X}\right|, \tag{6}$$

$$\vec{X}_1 = \vec{X}_{alpha} - \vec{A}_1.\vec{D}_{alpha}, \tag{7}$$

$$\vec{X}_2 = \vec{X}_{beta} - \vec{A}_1.\vec{D}_{beta}, \tag{8}$$

$$\vec{X}_3 = \vec{X}_{delta} - \vec{A}_1.\vec{D}_{delta}, \tag{9}$$

$$\vec{X}(t+1) = \frac{\vec{X}_1+\vec{X}_2+\vec{X}_3}{3}. \tag{10}$$

Feature selection is consider a binary space problem therefore the updated GWO position vector is forced to be binary utilizing the following formula:

$$\vec{X}(t+1) = \begin{cases} 1 & if\ sigmoid\left(\vec{X}(t+1)\right) \geq rand \\ 0 & otherwise \end{cases}. \tag{11}$$

The location of the ω wolf depending on the three best wolf's position is illustrated in Fig. 2. The ω wolf which requires position update is in the lower right corner.

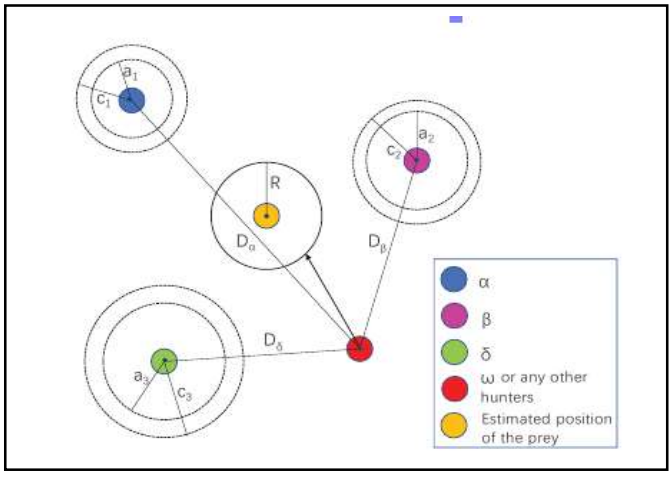

Figure 2. Position Updating.

Grey wolf optimization algorithm steps are illustrated in Fig. 3.

C. DECISION TREE

A decision tree (DCT) is a machine learning technique used to tackle classification issues. The primary goal of applying the DCT in this study work is to predict the target class utilizing decision rules derived from past data [10]. A fundamental classification and regression method is the decision tree. The classification of instances based on features may be performed by a model of a decision tree, which has structure of a tree [20]. Decision trees can be seen as conditional probability distributions or as a collection of if-then rules in class and attribute space. The CART (Regression and Classification Tree) technique is an alternative DCT building algorithm. It can solve both regression and classification tasks. This approach uses a new metric known as the Gini index to produce decision points for classification issues. The Gini index keeps the sum of squared probabilities for each class. It is depicted in the equation below [21, 22]:

$$Gini = 1\sum_{i=1}^{n}(p(i))^2, \tag{12}$$

where, p(i) is indicate label or class probability and n is the number of labels.

D. EVALUATION METRICS

Accuracy, recall, F-measure, and precision are utilized to evaluate the proposed method, and these criteria are computed as follows [23, 24]:

$$precision = \frac{tp}{tp+fp}, \tag{13}$$

$$recall = \frac{tp}{tp+fn}, \tag{14}$$

$$accuracy = \frac{tp+t}{tp+tn+fp+fn}. \tag{15}$$

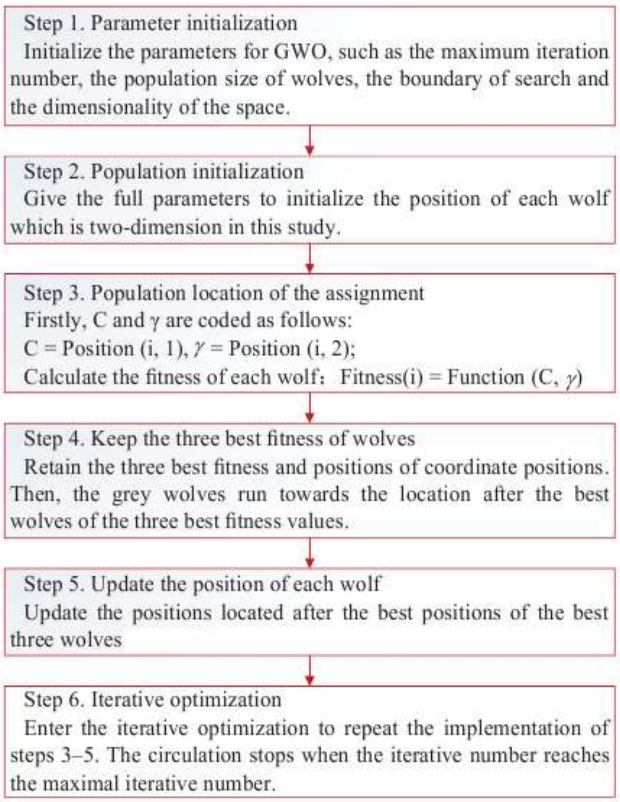

Figure 3. Diagram of the Regular Grey Wolf Optimizer Algorithm [18]

V. PROPOSED METHOD

The proposed method uses binary grey wolf optimization algorithm for feature selection and decision tree for classification process. The proposed method first preprocesses the dataset by replacing missing values by mean of the training set. Then dataset is normalized. Preprocessed data is used as input for the BGWO algorithm to produce the feature subset with the best fitness value. Proposed system is illustrated in Fig. 4.

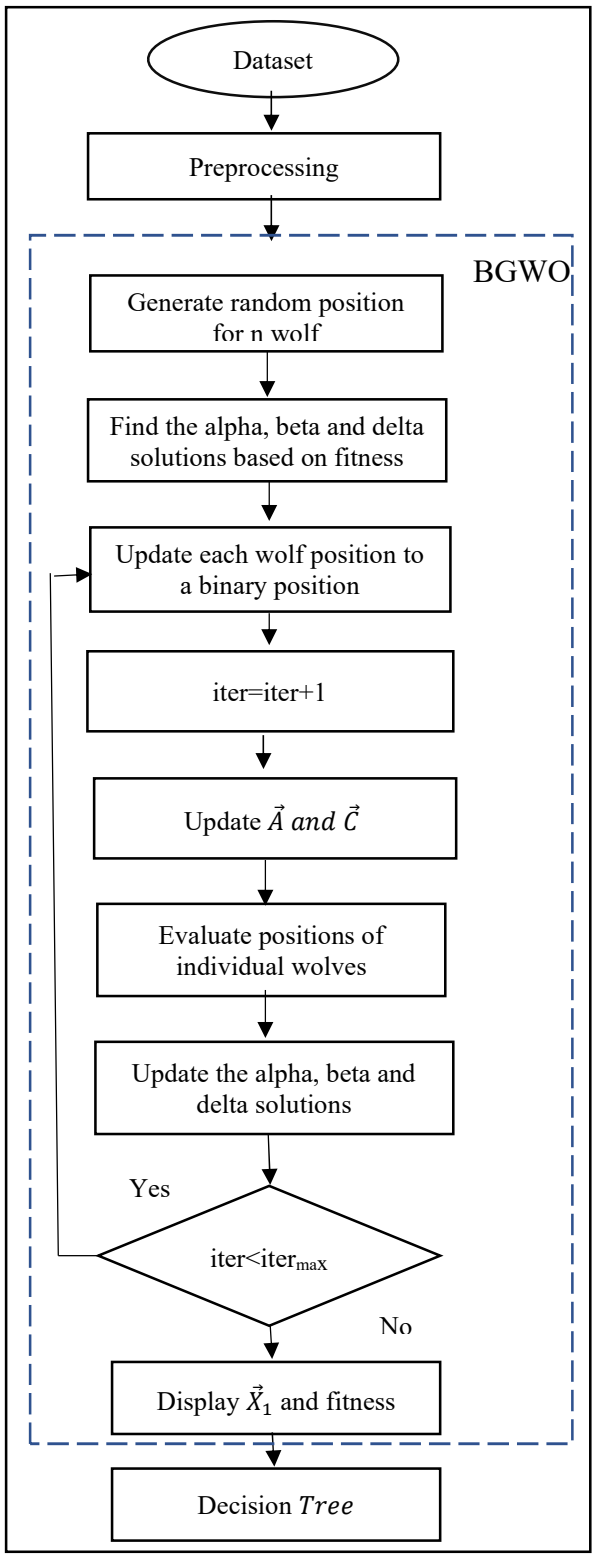

Figure 4. Block Diagram of the proposed method

Binary grey wolf optimization is a search strategy that may explore the attribute space adaptively to maximize the attribute evaluations criteria. In the search space, a single dimension presents an individual attribute, and so the wolf's location symbolizes a single attribute combination or solution. The proposed method is presented in Algorithm 1.

The nature of attribute selection is bi-objective. One objective is to maximize the accuracy of classification, and the other is to attain the minimum number of attributes. Therefore, the following equation is utilized as fitness function to consider both:

$$Fitness = \alpha * \big(1 - accuracy(DCT)\big) + (1 - \alpha)\frac{|S|}{|T|}. \quad (16)$$

Algorithm.1: Proposed System

Input: n: No. of wolf in pack

Output: $\vec{X}_1$: Best grey wolf position, f($\vec{X}_1$): value of best fitness.

Steps:

1. Initialize a random values ∈ [0,1] to an agent of n wolves positions.
2. Find alpha, mega, beta based on fitness function.
3. Evaluate the agent fitness using eq. (4-6).
4. while(t<maxiteration)
5. for each wolf in pack
6. Update $\vec{X}(t+1)$ using eq. (11).
7. end
8. update $\vec{X}_1, \vec{X}_2 and\ \vec{X}_3$ using eq.(7-9)
9. Evaluate the position of individual's wolves.
10. update $\vec{A}\ and\ \vec{C}$
11. end

VI. EXPERIMENTAL RESULTS

We implement the BGWO-DCT on Pima Indians Diabetes Database (PIDD) dataset [25]. The performance evaluation is done based on recall, accuracy, and precision metrics. The dataset samples are divided randomly into 70% for training and 30% for testing. Parameters of BGWO are the following: pack size is 30, and the number of iterations is 100.

The system consists of preprocessing, attribute selection, and classification processes. At preprocessing, the dataset is checked for missing values, replaced missing values with the mean of the training set, and normalized to prepare it for the feature selection process. The result showed that DCT when applied to PIDD dataset without feature selection, attained an accuracy of 77%. BGWO-DCT method is good combination for diabetes classification that an accuracy is increased from 77% with DCT to 83.11% with BGWO-DCT method. Table 2 illustrates the result of the proposed BGWO-DCT method.

Table 2. Result of Proposed Method

Method	Accuracy	Recall	Precision
DCT	77%	76%	77%
BGWO-DCT	83.11%	83%	82%

Comparison of the proposed BGWO-DCT and related work method is presented in Table 3. Fig. 5 illustrates a comparison between DCT results and the proposed method. Fig. 6 shows a comparison between the proposed method and the previous method. In comparison the proposed method achieved higher results than related work as illustrated in Table 3. In [8] and [9] authors only used the DM method for classification without feature selection. But in [6], the authors utilized PCA with logistic regression and achieved an accuracy of 79.56%, and the same accuracy was achieved with PSO and logistic regression. In [9], correlation and genetic algorithms for feature selection were utilized and an accuracy of 82.12% was achieved.

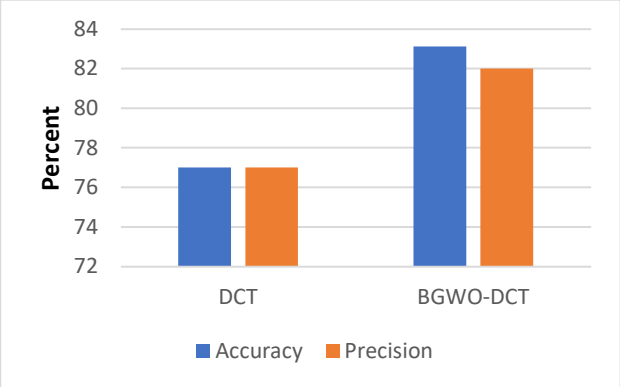

Figure 5. Results of the proposed method

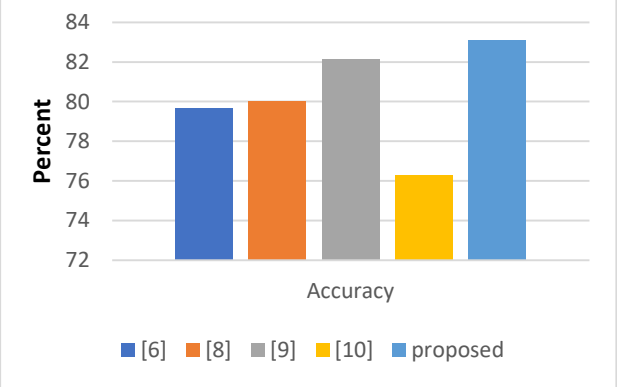

Figure 6. Comparison results of the proposed method

Table 3. Comparison of the Proposed Method with related work

Reference	Year	Method	Accuracy
[6]	2020	PCA and logistic regression PSO and logistic regression	79.56%
[8]	2017	KNN	80%
[9]	2021	PMSGD	82.12%
[10]	2018	NB	76.30%
Proposed		BGWO and DCT	83.11%

VII. CONCLUSIONS

An effective method of addressing diabetes classification problem was proposed. Several steps were implemented at preprocessing stage to enhance the classification accuracy. BGWO-DCT was utilized to solve the problem of feature selection in this work. To confirm the efficiency and the effectiveness of the proposed method PIDD dataset was employed. Accuracy, number of selected features, precision, and recall were used to assess the proposed model.'

The results demonstrate the superiority of the proposed method compared to a wide range of related work in terms of accuracy as shown in Table 3. The proposed BGWO-DCT achieved an accuracy of 83.11% with six features, whereas the classification accuracy of the DCT algorithm without feature selection attained an accuracy of 77%.

References

[1] American Diabetes Association, "Standards of medical care in diabetes – 2018 abridged for primary care providers," *Clin. Diabetes*, vol. 36, no. 1, pp. 14–37, 2018. https://doi.org/10.2337/cd17-0119.

[2] *IDF Diabetes Atlas*, International Diabetes Federation, 8th edition, 2017. https://diabetesatlas.org/upload/resources/previous/files/8/IDF_DA_8e-EN-final.pdf

[3] D. K. Choubey, S. Paul, V. K. Dhandhania, (2019). GA_NN: An Intelligent Classification System for Diabetes. In: Bansal, J., Das, K., Nagar, A., Deep, K., Ojha, A. (eds) Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing, 2019, vol 817, pp. 11-23. Springer, Singapore. https://doi.org/10.1007/978-981-13-1595-4_2.

[4] C. Mallika and S. Selvamuthukumaran, "A hybrid crow search and grey wolf optimization technique for enhanced medical data classification in diabetes diagnosis system," *Int. J. Comput. Intell. Syst.*, vol. 14, no. 1, pp. 1–18, 2021. https://doi.org/10.1007/s44196-021-00013-0.

[5] T. M. Le, T. M. Vo, T. A. N. N. Pham, S. O. N. Vu, and T. Dao, "A novel wrapper – based feature selection for early diabetes prediction enhanced with a metaheuristic," *IEEE Access*, vol. 9, pp. 7869–7884, 2021. https://doi.org/10.1109/ACCESS.2020.3047942.

[6] D. K. Choubey, P. Kumar, S. Tripathi, and S. Kumar, "Performance evaluation of classification methods with PCA and PSO for diabetes," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 9, no. 5, pp. 1–30, 2020. https://doi.org/10.1007/s13721-019-0210-8.

[7] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.*, vol. 9, pp. 1–10, 2018. https://doi.org/10.3389/fgene.2018.00515.

[8] S. Selvakumar, K. S. Kannan, and S. Gothainachiyar, "Prediction of diabetes diagnosis using classification based data mining techniques," *Int. J. Stat. Syst.*, vol. 12, no. 2, pp. 183–188, 2017.

[9] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," Multimed. Syst., vol. 28, pp. 1289-1307, 2021. https://doi.org/10.1007/s00530-021-00817-2.

[10] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018. https://doi.org/10.1016/j.procs.2018.05.122.

[11] R. Cheruku, D. R. Edla, V. Kuppili, and R. Dharavath, "A fuzzy rule miner integrating rough set feature selection and bat optimization for detection of diabetes disease," *Appl. Soft Comput.*, vol. 67, pp. 764–780, 2018. https://doi.org/10.1016/j.asoc.2017.06.032.

[12] A. M. Zeki, R. Taha, and S. Alshakrani, "Developing a predictive model for diabetes using data mining techniques," *Proceedings of the 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 2021, pp. 24–28. https://doi.org/10.1109/3ICT53449.2021.9582114.

[13] G. D. Kalyankar, S. R. Poojara, and N. V Dharwadkar, "Predictive analysis of diabetic patient data using machine learning and hadoop," *Proceedings of the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2017, pp. 619–624. https://doi.org/10.1109/I-SMAC.2017.8058253.

[14] A. B. García, "Brief update on diabetes for general practitioners," *Rev Esp Sanid Penit*, vol. 19, pp. 57–65, 2017.

[15] S. Watson, "Everything you need to know about diabetes," 2020. [Online]. Available at: https://www.healthline.com/health/diabetes.

[16] Mayo Clinic, "Diabetes," 2020. [Online]. Available at:

https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444.

[17] S. Mirjalili, S. Mohammad, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014. https://doi.org/10.1016/j.advengsoft.2013.12.007.

[18] X. Chen, J. Tuo, and Y. Wang, "A prediction method for blood glucose based on grey wolf optimization evolving kernel extreme learning machine," *Proceedings of the 2019 Chinese Control Conference (CCC)*, 2019, pp. 3000–3005. https://doi.org/10.23919/ChiCC.2019.8866210.

[19] Q. Li *et al.*, "An enhanced grey wolf optimization based machine for medical diagnosis," *Comput. Math. Methods Med.*, vol. 2017, pp. 1–15, 2017. https://doi.org/10.1155/2017/9512741.

[20] J. R. Quinlan, "Generating production rules from decision trees," *Proceedings of the 10th International Joint Conference on Artificial Intelligence IJCAI'87*, 1987, vol. 87, pp. 304–307.

[21] E. Z. Aziza, L. M. El Amine, M. Mohamed, and B. Abdelhafid, "Decision tree CART algorithm for diabetic retinopathy classification," *Proceedings of the 6th International Conference on Image and Signal Processing and Their Applications (ISPA)*, 2019, pp. 1–5. https://doi.org/10.1109/ISPA48434.2019.8966905.

[22] B. S. Babu, A. Suneetha, G. C. Babu, Y. J. N. Kumar, and G. Karuna, "Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network," *Period. Eng. Nat. Sci.*, vol. 6, no. 1, pp. 229–240, 2018. https://doi.org/10.21533/pen.v6i1.286.

[23] N. A. Saeed and Z. T. M. Al-Ta'i, "Feature selection using hybrid dragonfly algorithm in a heart disease predication system," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 2862–2867, 2019. https://doi.org/10.35940/ijeat.F8786.088619.

[24] N. A. Saeed and Z. T. M. Al-Ta'i, "Heart disease prediction system using optimization techniques," in *New Trends in Information and Communications Technology Applications*, 2020, pp. 167–177. https://doi.org/10.1007/978-3-030-55340-1_12.

[25] "UCI Machine Learning Repository: Pima Indians Diabetes Database," [Online]. Available at: https://archive.ics.uci.edu/ml/index.php.

LAYLA A. AL.HAK, M. S. C. Computer Science, Department of Computer Science, College of Science, University of Diyala, Iraq.