A Hybrid Feature Selection Method for Improve The Accuracy of Medical Classification Process

Maria Mohammad Yousef

***Abstract**: Generally, medical dataset classification has become one of the biggest problems in data mining research. Every database has a given number of features but it is observed that some of these features can be redundant and can be harmful as well as disrupt the process of classification and this problem is known as a high dimensionality problem. Dimensionality reduction in data preprocessing is critical for increasing the performance of machine learning algorithms. Besides the contribution of feature subset selection in dimensionality reduction gives a significant improvement in classification accuracy. In this paper, we proposed a new hybrid feature selection approach based on (GA assisted by KNN) to deal with issues of high dimensionality in biomedical data classification. The proposed method first applies the combination between GA and KNN for feature selection to find the optimal subset of features where the classification accuracy of the k-Nearest Neighbor (kNN) method is used as the fitness function for GA. After selecting the best-suggested subset of features, Support Vector Machine (SVM) are used as the classifiers. The proposed method experiments on five medical datasets of the UCI Machine Learning Repository. It is noted that the suggested technique performs admirably on these databases, achieving higher classification accuracy while using fewer features.*

Keywords: Dimensionality Problem, Feature Selection, classification, Genetic Algorithm.

I. INTRODUCTION

Data mining, also known as Knowledge Discovery in Databases (KDD), is one of the most crucial topics in the field of research that is used for automatically discover, evaluate, interpret, and find new hidden patterns from massive amounts of real data sets kept in repositories [1]. There are many tasks of data mining techniques including classification [2], clustering [3], association mining rule[3], and pattern recognition [4]. The classification task is also known as supervised learning, and it involves learning from training data sets to develop the classifier. In training data, there are frequently a huge number of irrelevant or redundant features, which has a significant impact on subsequent classification accuracy and machine learning efficiency Therefore, It is observed that when the features of the training data sets exceed a certain range of the sample space, the accuracy of the classifier will decrease and this is known as a high dimensionality problem.

High-dimensional data space can raise the computational cost in addition to reducing the classifier prediction accuracy. So, dimensionality reduction for the classification process plays a vital role in improving the performance of both machine learning and pattern recognition. The dimensionality reduction can be accomplished in two ways: feature selection and feature extraction and [5], [6]. As is mentioned in [5], feature selection is also defined as feature subset selection.

The fundamental purpose of feature selection in the classification process is to eliminate irrelevant and redundant features while retaining the most important information from the original data. To put it another way, the idea of feature selection is to select an optimal features subset from the original features set.

As a crucial step of knowledge discovery technology, Appropriate feature selection can significantly increase the prediction algorithm's computing speed, facilitates data visualization and improves the understanding of the computational models, enhances classification accuracy, reduces the requirements of measurement and storage and thus minimizes the cost in database storage and management, and avoids overfitting.

In this paper we produced a new hybrid approach to select optimal subset of features by combination between genetic algorithm and KNN the best subset of features which has high classification accuracy on given database is adopted. In this study the model training and evaluating using 5 dataset collected from UCI machine learning repository. The rest of the paper is organized as follows: Section II presents the review of dimensionality reduction techniques to analyze how effective these techniques can be used to achieve high performance of learning algorithms that ultimately improve the predictive accuracy of the classifier. Section III presents a brief explanation of the algorithms used. Section IV explains the methodology of the research. Where Section V shows the experimental result and Section VI concludes the paper.

II. RELATED WORK

Feature selection is a mature area of research. Several feature selection algorithms have been proposed in recent years as many studies have examined the effectiveness of different feature selection algorithms based on different data sets. In this section, we will provide a brief overview of the feature selection methods used in medical dataset. In [7], the authors suggested the NB-SKDR, a new heart disease prediction model based on the Naive Bayes

Manuscript received on November 13, 2021.
Revised Manuscript received on November 20, 2021.
Manuscript published on November 30, 2021.

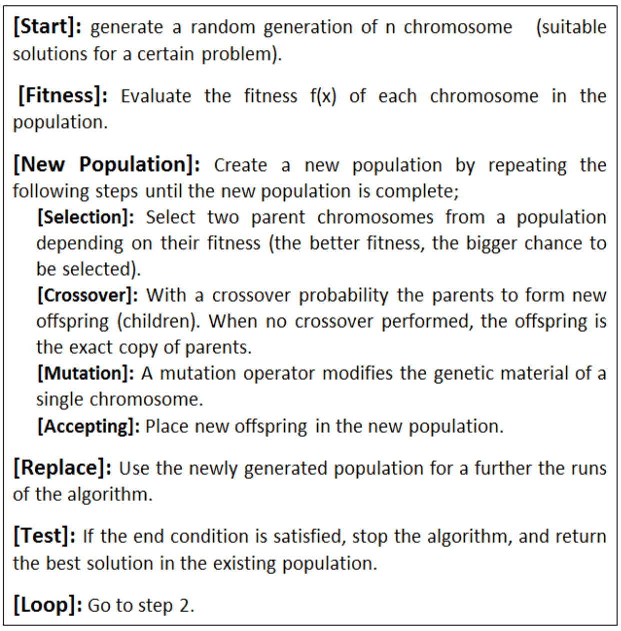
* Correspondence Author
Maria Mohammad Yousef*, Department of Computer Science, Al al-Bayt University, Jordan

algorithm (NB) as a feature selection method and multiple machine learning techniques as classifiers, including Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest. The main objective of this study is to identify the best subset of features for classification and optimize the accuracy of the heart disease prediction system based on the Cleveland Heart Disease dataset which represents the historical medical files and contains 303 records with 13 attributes. The (NB) method is used to find the most important features by calculating the dependent probability between each pair of features using the Bayes rule. This method has been able to reduce features from 13 to 6. The performance of the classifiers was examined, and their results reached 98% of accuracy when using the SVM classifier, while DT and KNN achieved 95%, 97% of accuracy respectively. In [8], The authors presented an efficient feature selection approach to improve the efficiency and reliability of the chronic kidney disease diagnosis system by combining correlation coefficient and recursive feature elimination to reduce the number of features by eliminating irrelevant and useless features .the performance of the Decision Tree (SVM), Naive Bayes (NB), and Random Forest (RF) classifier have been compared based on its accuracy, precision, Recall and F-Score for kidney prediction. In this study, the dataset used includes the 400 cases in addition to 24 features presented in different formats; the experimental results revealed that the performance of the RF classifier is better than DT, and NB its achieved 100% of accuracy. [9] provided a new approach to select a reduced number of features in databases. This research method applies a binary coded genetic algorithm to select a small subset of features. The relevance of these features is assessed using the Nave Bayes (NB) classification algorithm. where the best-reduced subset of features has high classification accuracy. In this study, the experiments were performed on eight medical databases namely Ionosphere, Australian, Wine, Sonar, German, Heart, Wisconsin Diagnosis breast cancer, (WDBC) datasets, Wisconsin Prognostic Breast Cancer (WPBC). The results showed that the use of the GA algorithm to select the best subset of features led to an increase in the accuracy of classification in medical databases.

III. PRELIMINARY MATERIAL

A. Genetic Algorithm

The genetic algorithm have been developed by John Holland in 1975 [10]. It's an evolutionary-inspired heuristic model for solving a specific optimization problem. Moreover, GA is a Randomized search algorithm based on the mechanics of natural genetics and the notion of gene evolution. The algorithm begins by creating a population of potential solutions that are encoded as a string named chromosomes. Each solution has a fitness value, which is used to determine which parents should be utilized for reproduction (survival of the fittest). The new generation of a chromosome is produced by employing some generative operators such as selection (To generate the mating pool, the natural genetic selection was used), crossover (information sharing between parents), and mutation (a sudden small modification in a parent) on selected parents. Thus the quality of the population is improved as the number of generations increases. The Procedure of GA is repeated until a certain condition is satisfied or the solution converges to the target of a problem. This algorithm has been proven theoretically and experimentally that it can search for the optimal solution in complex space. Figure 1 illustrates the steps of GA.



[Start]: generate a random generation of n chromosome (suitable solutions for a certain problem).

[Fitness]: Evaluate the fitness f(x) of each chromosome in the population.

[New Population]: Create a new population by repeating the following steps until the new population is complete;

- **[Selection]:** Select two parent chromosomes from a population depending on their fitness (the better fitness, the bigger chance to be selected).
- **[Crossover]:** With a crossover probability the parents to form new offspring (children). When no crossover performed, the offspring is the exact copy of parents.
- **[Mutation]:** A mutation operator modifies the genetic material of a single chromosome.
- **[Accepting]:** Place new offspring in the new population.

[Replace]: Use the newly generated population for a further the runs of the algorithm.

[Test]: If the end condition is satisfied, stop the algorithm, and return the best solution in the existing population.

[Loop]: Go to step 2.

Figure 1: The Steps of GA.

Genetic algorithm has primary operators:

- **Coding of the population:** The process starts with a set of individuals which is named a Population. An individual is described by a set of parameters (variables) known as Genes. Genes are joined into a string to create a Chromosome (solution) that represents feature subsets and is randomly initialized at the start. The length of each chromosome equals the length of the feature set. In a genetic algorithm, the set of genes of an individual is represented using a string, usually binary values are used (1, 0) [11].
- **Fitness function**: In the coding step, each solution is generally represented as a string of binary numbers, referred to as a chromosome. We have to test these solutions in order to find the best set of solutions to solve a given problem. Each solution, therefore, needs to be awarded a score, to indicate how close it came to meeting the overall specification of the optimum solution (the ability of an individual to compete with other individuals). This score is generated by applying the fitness function (also known as the Evaluation Function) to the test, or results obtained from the tested solution [12].
- **Crossover:** One of the most important genetic operators is a crossover, which combines (mates) two chromosomes (parents) to create a new chromosome (offspring). The idea behind crossover is that the new chromosomes may be better than both parents if they take the best characteristics from each of the parents. Crossover occurs during evolution according to a user-definable crossover probability.

Single-point crossover, two-point crossover, and uniform crossover are three popular crossover approaches [13].

- **Mutation:** is a genetic operator that randomly alters one or more gene values in a chromosome from its initial state and protects against converging too rapidly to a local optimum. This may result in the addition of entirely new gene values to the gene pool. With these new gene values, the Genetic Algorithm may be able to arrive at a better solution than was previously possible. Without mutation, all the combinations that we would ever possibly reach during the successive generations would be already in the initial pool. For binary encoding a few randomly chosen bits are changed from 1 to 0 or 0 to 1.

B. K-Nearest Neighbor Algorithm

The K-nearest neighbor algorithm is a classification approach based on case similarity. The kNN is an instance-based classifier that operates on the assumption that the classification of unknown cases may be determined by comparing the unknown to the known cases using some distance or similarity measurement. Those close to others are called "neighbors". When a case is new, its distance from each of the cases in the model is calculated. Applying this classification specifies the case as being the nearest neighbor, which is the most similar. Therefore, it puts the case into the group that contains the nearest neighbors. the two instances that are separated by a large distance in the instance space described by the appropriate distance function are less likely to belong to the same class as two instances that are near together.

To employ the algorithm, we require the following details (algorithm input):

1. A set of labeled stored records for evaluating the class of a test object (training dataset);
2. A distance scale (metric) to estimate the similarity between objects. that can be used to calculate the closeness normally of objects;
3. The k value is the number of nearest neighbors (records) belonging to the training dataset, based on which we will accomplish the classification of a new object.

IV. THE PROPOSED APPROACH

In this study, we proposed a hybrid feature selection approach by using GA assisted by a kNN to improve the classification accuracy of five benchmarked medical datasets described in Table 1. Whereas the primary achievement of this work was to optimize the classification process and to identify an optimal subset of features. Figure 2 depicts the proposed model methodology. The proposed model is composed of two main phases which include: feature selection, and classification:

A. Feature Selection Based ON GA and KNN

To select an optimal subset of features, we employed a hybrid approach that included GA and kNN. In this manner, The GA is used in the search to obtain feature subsets, where each subset corresponds to be a chromosome (in the evolutionary context). Binary chromosomes are generated randomly to form an initial population in the genetic algorithm. Then, the fitness value is calculated using the fitness function for each chromosome in the population; the kNN method has been used for this purpose. The GA parameters are listed in Table 1.

Table 1. GA parameters and their values

Parameters	**Values**
Chromosome size	24, 20, 23, 34, and 32
Population Type	Bit String
Population Size	50
Maximum no. of Generation	100
Selection Method	Roulette Wheel
Selection Rate	50%
Crossover Method	Single point
Crossover Probability	0.5
Mutation Method	Flip bit
Mutation Probability	0.1
Distance Measure in kNN	Euclidean distance
k-value in kNN	5

***1. Chromosome Representation*:** The chromosome should include information about the solution which it represents. In this step, we used the bit string type to represent the genes of the chromosome (population) 0 or 1. At the start, the population of 50 chromosomes that represent feature subsets is randomly initialized, and random values for gene position are created, the genes are considered when the value in its position is greater than 0.5, otherwise, it is ignored. The length of each chromosome equals the length of the feature set that equals 24, 20, 23, 34, and 32 features for each dataset. The GA is configured to have 50 chromosomes and was run for 100 generations in each trial.

2. The Fitness function: Now, after a set of chromosomes has been generated in the initial population (which denotes feature subsets), the initial population chromosomes were used as input data for the kNN to realize the output aim of assessing these chromosomes in order to choose the best subset of features. Initially, the kNN method starts by searching feature values equal to 1 inside the population (i.e. GA has selected it initially). Then, to search for the nearest neighbors the exhaustive searcher has been used, and the distance type between the features and the classes is the Euclidean distance. Afterward, the kNN classification object is searched for 3 nearest neighbors. The accuracy of the kNN classifier is used as the fitness function for GA. The fitness function fitness(x) is defined as in eq 1.

$$fitness(x) = Accuracy(x) \qquad (1)$$

Accuracy(x) is the test accuracy of testing data x in the kNN classifier which is built with the feature subset selection of training data. The classification accuracy of kNN is given by eq 2 [14].

$$Accuracy(x)= (c\ /t)*100 \qquad (2)$$

Where:

c c- denotes the number of samples correctly classified in test data using the kNN algorithm.

t- represents the total number of samples in the test data.

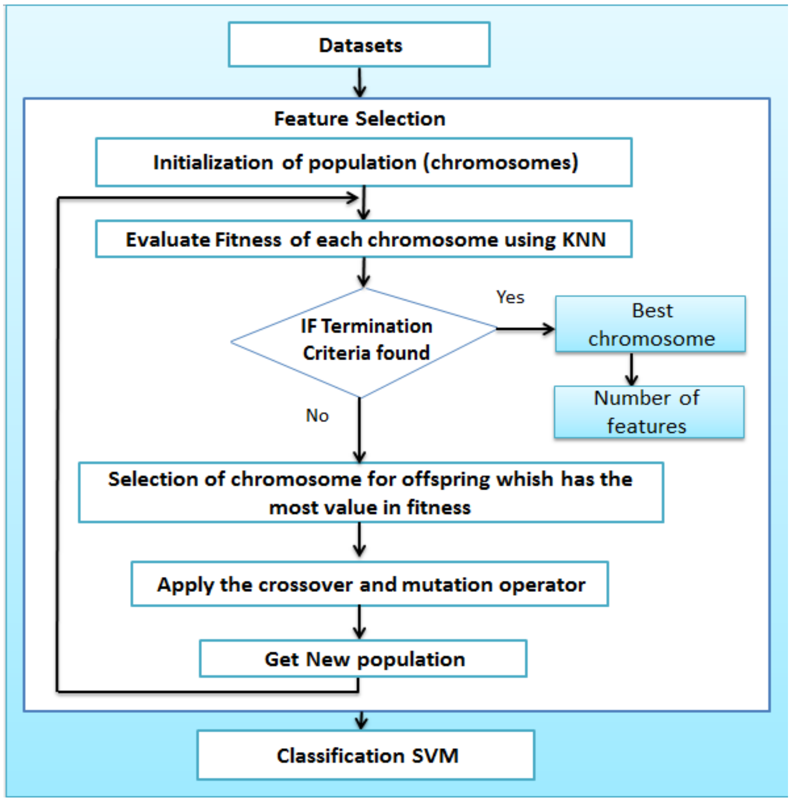

Datasets
Feature Selection
Initialization of population (chromosomes)
Evaluate Fitness of each chromosome using KNN
IF Termination Criteria found
Yes
Best chromosome
Number of features
No
Selection of chromosome for offspring whish has the most value in fitness
Apply the crossover and mutation operator
Get New population
Classification SVM

Figure 2: The Proposed Model Methodology

B. Classification Based On SVM

In the last stage, the relationship between the attributes (a new subset of features) is taken as input for this stage, the input data is divided into a training set and test set. The algorithm of the support vector machine will be applied on the selected features for the final classification. The Support Vector Machine models are defined as finite-dimensional vector space in which each dimension represents a 'feature" of a particular object [15]. It has been shown to be an effective approach in high-dimensional space problems. Due to its computational efficiency on large datasets. The execution of the proposed approach will be investigated as far as specific parameters like accuracy.

V. EXPERIMNTEAL RESULT:

The programming tool used to perform the experiments in this study is MATLAB, which provides special libraries to train and test machine learning models with different parameters settings. Also, GA MATLAB code is used for this study. The proposed method is applied over the five medical datasets as described in Table 2.

A. Dataset

This study is experimented based on five different medical datasets obtained from UCI Machine Learning and available at: (https://archive.ics.uci.edu/ml/datasets.html). Table 1 described the details of various medical datasets, including the number of instances, features, and classes. Both the training and testing datasets are created at random, The 75 % of the dataset is used for training and 25 % is used in testing the proposed method.

Table 2. Description of Medical Datasets.

Dataset	Number of Rows	Number of features	Number of classes
Heart Disease (SPECTF)	267	44	2
Diabetic Retinopathy Debrecen	1151	20	2
Cardiotocography (CTGs)	2126	23	3
Wisconsin Breast Cancer (Prognostic)	198	34	2
Wisconsin Breast Cancer (Diagnostic)	569	32	2

The Wisconsin Breast Cancer (Diagnostic) dataset is published by the University of Wisconsin. It contains 569 cases with 32 features that can be used to determine whether a growth is benign or cancerous [16]. The Wisconsin Breast Cancer (Prognostic) dataset was collected from the University of Wisconsin. which includes 198 examples with 20 features that can be utilized to identify recurrent and nonrecurring events [16]. Diabetic Retinopathy is a dataset obtained from the University of Debrecen that comprises 1151 cases with 20 variables that can be used to determine whether the patient has diabetic retinopathy or not [17]. Cardiotocography (CTGs), this dataset was generated by Diogo Ayres-de-campos at the University of Porto. It has 2126 occurrences with 23 characteristics that can be used to predict fetal conditions [18]. Heart Disease (SPECTF), this dataset is based on data from the University of Colorado.

It has 45 features with 267 instances that are utilized to determine whether patients are suffering from the disease or not [19].

B. Result and Analysis

In this section, the proposed hybrid feature selection method (GA and KNN) was evaluated in its ability to handle the high dimensional problem and its effectiveness in increasing the classification accuracy when selecting the optimal training parameters. In the first experiment, the SVM algorithm relied on all the features in the different databases during the classification process, while in the second experiment, the SVM algorithm based on the best subset produced from the feature selection step and used it as input to perform the classification process.
Table 3 displays the results of the experiments. Table II has five columns: the first column lists the dataset name, the second column lists the number of optimal features (selected by GA with KNN), the third column lists the total number of features, the fourth column lists the SVM accuracy (first experiment), and the fifth column lists the proposed method's accuracy (second experiment).

Table 3. Experiment Results of Featured Selection (GA with KNN)

Dataset	Total Number of Features	Number of Optimal Features	SVM Accuracy First Experiment	Proposed Method Accuracy Second Experiment
Heart Disease (SPECTF)	44	10	77.5%	87.5%
Diabetic Retinopathy Debrecen	20	9	61.1%	71.6%
Cardiotocog raphy (CTGs)	23	9	90.9%	98.5%
Wisconsin Breast Cancer (Prognostic)	34	12	78.7%	86.4%
Wisconsin Breast Cancer (Diagnostic)	32	5	94.1%	99.2%

The experimental findings in Table 2 showed that the suggested feature selection approach may improve the accuracy of the five medical datasets by 5% to 10% increase in comparison with the SVM algorithm without optimization and features selection. The highest optimized performance was received from the classification of the Diabetic Retinopathy dataset with an increase of 10.5% of 61.1% with the most optimal 9 features. Meanwhile, the classification of the Wisconsin Breast Cancer (Diagnostic) dataset showed the lowest improvement, with accuracy increasing by only 5.1% from 94.1% when relying on the best 5 features to make the classification. Furthermore, selecting the 12 best out of 34 features leads to improved performance on the Wisconsin Breast Cancer (prognostic) dataset by 7.7% from 78.7%. The Cardiotocography dataset also increased by 7.6% from the original 90.91% when reducing the number of features to 9. As in the Heart Disease (SPECTF) dataset, classification accuracy increased by 10% from 77.5% when selecting the top 10 features.

VI. CONCLUSION:

In this research, a novel feature selection approach has been presented to improve the classification performance of five benchmarked medical datasets. Whereas the main achievement of this research was to handle the high dimensionality problem by finding an optimal subset of features for the classification process. The proposed method consists of two stages which include: feature selection, and classification. In the feature selection, the hybrid method has been adopted to reduce the dimensionality problem and to select an optimal subset of features based on the GA assisted by the kNN. The GA is used to search for the optimal feature subset by calculating the fitness value for each chromosome through the accuracy of kNN. The SVM algorithm was also applied to the data sets for the purpose of classification. The results show that the classification accuracy for each database increased between 5% to 10% when using the feature set obtained from combining GA and KNN.

REFERENCES

1. N.Tomasevic, N. Gvozdenovic, S. Vranes," An overview and comparison of supervised data mining techniques for student exam performance prediction.", Computers and Education, Vol.143, 2020, pp.103676.
2. A.Tharwat, "Classification assessment methods", Applied Computing and Informatics", Vol. 17, No.1, 2018, pp.168–192.
3. A. K. Mann, & N. Kaur, "Review paper on clustering techniques". Global Journal of Computer Science and Technology, Vol.13, No. 5, 2013, pp. 44-48.
4. M. Paolanti, & E. Frontoni, "Multidisciplinary Pattern Recognition applications: A review.", Computer Science Review, Vol. 37, 2020, pp. 100276.
5. G. Chandrashekar, & F. Sahin, "A survey on feature selection methods.", Computers and Electrical Engineering, Vol. 40, No.1, 2014, pp.16–28.
6. S. Khalid, T. Khalil, & S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning.", Proceedings of 2014 Science and Information Conference, SAI 2014, October 2016, pp.372–378.
7. M. Yousef, & P. K. Batiha, "Heart Disease Prediction Model Using Naïve Bayes Algorithm and Machine Learning Techniques.", International Journal of Engineering & Technology, Vol. 10, No.1, 2021, pp.46–56.
8. M. Yousef, "Prediction of chronic kidney disease using different classification algorithms.", Journal of Xi'an Shiyou University, Natural Science Edition, Vol.24, No.10, 2021, pp.453–462.
9. A. Saxena, & M. M. Shrivas, "Leave one out cross validated Hybrid Model of Genetic Algorithm and Naïve Bayes for Classification Problem.", Vol. 6, No.3, 2016, pp.107–114.
10. F. Vericat, C. O. Stoico, C. M. Carlevaro, & D. G. Renzi, "Genetic algorithm", Interdisciplinary Sciences: Computational Life Sciences, Vol. 3, No. 4, 2011, pp.283–289.
11. Y. He, & C. W. Hui, " A binary coding genetic algorithm for multi-purpose process scheduling: A case study.", Chemical Engineering Science, Vol. 65, No.16, 2010, pp. 4816–4828.
12. F. Alabsi, & R. Naoum, " Fitness Function for Genetic Algorithm used in Intrusion Detection System.", International Journal of Applied Science and Technology, Vol. 2, No.4, 2012, pp. 129–134.
13. D. Vrajitoru, "Crossover improvement for the genetic algorithm in information retrieval.", Information Processing and Management, Vol. 34, No.4, 1998, pp. 405–415.
14. P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, & H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview.", Bioinformatics, Vol. 16, No.5, 2000, pp. 412-424.
15. C. Cotrtes, & V. Vapnik, "Support-vector networks.", Machine learning, Vol. 20, No.3, 1995, pp. 273-297.

16. O. L. Mangasarian, W. N. Street, & W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming.", Operations Research, Vol. 4, No.3, 1995, pp. 570-577. .
17. B. Antal, & A. Hajdu, "An ensemble-based system for automatic screening of diabetic retinopathy.", Knowledge-Based Systems, Vol. 60, 2014, pp. 20-27.
18. D. Ayres-de-campos et al., "SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms.", The Journal of Maternal-Fetal Medicine, Vol. 9, 2000, pp. 311-318.
19. L. A. Kurgan et al., " Knowledge discovery approach to automated Cardiac SPECT Diagnosis.", ArtificialIntelligence in Medicine, Vol. 23, 2001, pp. 149-169.

AUTHORS PROFILE

Maria Mohammad Yousef, received a Master degree in computer science from Al al-Bayt University, Jordan, in 2020 and graduated with first class honors. She holds a BA degree in Computer Science from Al al-Bayt University, Jordan, in 2016 and graduated with 1st class Honor. Recently, she is working as a lecturer in special organization. Her field of research includes Data mining, Big Data, Healthcare, Machine learning.