A MODIFIED MODEL BASED ON FLOWER POLLINATION ALGORITHM AND K-NEAREST NEIGHBOR FOR DIAGNOSING DISEASES

MEHDI ZEKRIYAPANAH GASHTI

Department of Computer Engineering, Payame Noor University, Tehran, Iran

Corresponding author: gashti@pnu.ac.ir

(Received: 17th July 2017; Accepted28th Feb 2018; Published on-line: 1st June 2018)

https://doi.org/10.31436/iiumej.v19i1.854

ABSTRACT: Exponential growth of medical data and recorded resources from patients with different diseases can be exploited to establish an optimal association between disease symptoms and diagnosis. The main issue in diagnosis is the variability of the features that can be attributed for particular diseases, since some of these features are not essential for the diagnosis and may even lead to a delay in diagnosis. For instance, diabetes, hepatitis, breast cancer, and heart disease, that express multitudes of clinical manifestations as symptoms, are among the diseases with higher morbidity rate. Timely diagnosis of such diseases can play a critical role in decreasing their effect on patients' quality of life and on the costs of their treatment. Thanks to the large data set available, computer aided diagnosis can be an advanced option for early diagnosis of the diseases. In this paper, using a Flower Pollination Algorithm (FPA) and K-Nearest Neighbor (KNN), a new method is suggested for diagnosis. The modified model can diagnose diseases more accurately by reducing the number of features. The main purpose of the modified model is that the Feature Selection (FS) should be done by FPA and data classification should be performed using KNN. The results showed higher efficiency of the modified model on diagnosis of diabetes, hepatitis, breast cancer, and heart diseases compared to the KNN models.

ABSTRAK: Pertumbuhan eksponen dalam data perubatan dan sumber direkodkan daripada pesakit dengan penyakit berbeza boleh disalah guna bagi membentuk kebersamaan optimum antara simptom penyakit dan mengenal pasti gejala penyakit (diagnosis). Isu utama dalam diagnosis adalah kepelbagaian ciri yang dimiliki pada penyakit tertentu, sementara ciri-ciri ini tidak penting untuk didiagnosis dan boleh mengarah kepada penangguhan dalam diagnosis. Sebagai contoh, penyakit kencing manis, radang hati, barah payudara dan penyakit jantung, menunjukkan banyak klinikal simptom jelas dan merupakan penyakit tertinggi berlaku dalam masyarakat. Diagnosis tepat pada penyakit tersebut boleh memainkan peranan penting dalam mengurangkan kesan kualiti hidup dan kos rawatan pesakit. Terima kasih kepada set data yang banyak, diagnosis dengan bantuan komputer boleh menjadi pilihan maju menuju ke arah diagnosis awal kepada penyakit. Kertas ini menggunakan Algoritma Flower Pollination (FPA) dan K-Nearest Neighbor (KNN), iaitu kaedah baru dicadangkan bagi diagnosis. Model yang diubah suai boleh mendiagnosis penyakit lebih tepat dengan mengurangkan bilangan ciri-ciri. Tujuan utama model yang diubah suai ini adalah bagi Pemilihan Ciri (FS) perlu dilakukan menggunakan FPA and pengkhususan data perlu dijalankan menggunakan KNN. Keputusan menunjukkan model yang diubah suai lebih cekap dalam mendiagnosis penyakit kencing manis, radang hati, barah payudara dan penyakit jantung berbanding model KNN.

KEYWORDS: *flower pollination algorithm; K-nearest neighbor; data classification;*

feature selection; diagnosis diseases

1. INTRODUCTION

Chronic diseases such as diabetes, hepatitis, breast cancer and heart disease are the main cause of death in many countries [1, 2]. Early diagnosis and treatment of these diseases can provide a tremendous effect on improving patients' quality of life and reducing the costs of their treatment. An effective treatment can be performed either based on the direct and subjective diagnosis by a clinical care team and their advice or based on prediction-based diagnosis. In the prediction-based method, retrospectively collected data from patients and healthy subjects are exploited to determine the probability of particular disease incidence in each person based on his/her medical data. Appropriate computer-based methods can handle huge medical data sets in a short time period in order to predict disease incidence based on collected data.

Different symptoms of the particular disorder can be considered based on their importance and occurrence. For instance, diabetes, a metabolic disease due to the lack of insulin production or to resistivity to insulin, is characterized by high blood glucose levels [2]. Hepatitis, on the other hand, is known by inflammation of the liver parenchyma and can be resulted from several etiologies; some of them are contagious and others are not. Among the factors creating hepatitis are excess in alcohol consumption, the effects of some medications, and infection with bacteria or viruses [3]. Heart disease is mainly manifested by partial or complete arterial occlusion. Arterial disorders consequently affect the availability of the blood supply and therefore nutrients and oxygen to different organs including the heart itself [4]. One of the most effective strategies to cope with cardiovascular disorders is to identify the risk factors that play a critical role in the development of such diseases. Breast cancer is also one of the most common diseases in modern societies, and a combination of the genetic and environmental factors can affect its incidence [5]. Its early diagnosis, by targeting the risk factors involved, plays a critical role on efficiency of the treatment.

As the number of features for a particular disease increases, the disease's diagnosis and prognosis become increasingly challenging, even for a well-qualified medical professionals. To cope with this issue, in recent decades, computer-based diagnosis tools have been developed to assist physicians. Computer-based tools offer several advantages including higher speed, precision, and lack of fatigue or condition-dependent decisions [6]. Analysis and modeling tools such as artificial intelligence algorithms possess great potential to deal with huge data sets collected from patients, a feature that can significantly improve medical decisions including diagnosis and treatment selection.

The purpose of this paper is to offer a new way of diagnosis for diseases such as diabetes, hepatitis, breast cancer, and cardiovascular disorders. In the modified model, FPA [7] helps feature extraction and KNN [8] is used for data classification. The basic FS problem is an optimization problem with a performance measure for each subset of features to measure its ability to classify the samples. The aim of modified model is to reduce the dimensions and characteristics of the sample space, and ultimately, improve the accuracy of diagnosis and classification of patients. Final results showed that the modified model has more diagnosis accuracy compared with KNN [8]. Classification, belonging to supervised learning, is one of the most popular data mining techniques [9]. In supervised learning, each instance in the training data is labeled with a class. The task of classification can be divided into two parts: training and testing. Testing uses a classifier that is trained to assign a class label to a new unlabeled instance.

The overall structure of this paper is organized as follows: In Section 2, we explain the related models that have been developed for diagnosis of diabetes, hepatitis, breast cancer, and heart disease. In Section 3, the modified model is explained. The results of the modified model are presented in Section 4. Finally, our results are compared with previous models in Section 5.

2. RELATED WORK

Previous methods that have been tried on disease diagnosis include Probabilistic Neural Networks (PNN) and Fuzzy Classification Systems based on Ant Colony Optimization (FCS-ANTMINER). PNN is a model of Artificial Neural Networks (ANNs) that has been used on 768 samples and 8 features for diabetes diagnosis using [10] a Pima dataset. PNN often learns more quickly than many ANN models, such as back propagation networks, and has been exploited for different applications. Different neurons are used for testing and training in MATLAB. The results showed that a PNN model with 82.37% accuracy has better prediction power compared to other models. Fuzzy Classification System based on an Ant Colony Optimization (FCS-ANTMINER) [11] is a hybrid of the Ant Colony Optimization (ACO) and fuzzy is implemented and tested on a Pima dataset for the diagnosis of diabetes. An ACO-based classification system extracts a set of fuzzy rules for diagnosis of diabetes, named FCS-ANTMINER. The ACO model is used for rules and the fuzzy for data classification. The results showed that the accuracy of the FCS-ANTMINER model is 84.32%.

A case-based reasoning (CBR) [12] model is a fuzzy-ontology hybrid that was implanted and tested on 2,640 diabetic patients. The system was implemented in six modules: case source preparation, case-based ontology engineering, terminology server, fuzzy case-based ontology population, case retrieval engine, and case query parser. The ontology model was used for communication between the diseases and attributes. Fuzzy classification was used to establish rules and diagnosis. The results showed that the accuracy of the fuzzy-ontology model was 97.67%. The Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System (LDA-ANFIS) [13] model is a hybrid of neural networks and fuzzy that was implemented and tested on 768 subjects for diabetes diagnosis. In LDA-ANFIS mode, fuzzy rules have been used to derive and obtain a precise amount. Results showed that the recognition accuracy in LDA-ANFIS model was 84.61%.

The hybrid K-Means-Genetic Algorithm-Support Vector Machine (K-Means-GA-SVM) [14] model was implemented and tested to detect and predict diabetes on 768 people. The K-Means model was used for data clustering and preparing data for SVM classification. The basic concept of SVM is finding an optimal separating hyper-plane to classify the separable data by maximizing the margin between points from each class. The points lying on the boundaries are called support vectors. As well, GA has been used for selecting the appropriate features. The results showed that the accuracy of K-Means-GA-SVM model was 98.82%. The Modified Artificial Bee Colony (MABC) [15] Model is a hybrid of the Artificial Bee Colony (ABC) and fuzzy that was suggested for classification of diabetic patient data. ABC was used for FS and fuzzy for the formation of healthy/unhealthy laws. Results of more than 768 data have shown that the MABC model was more accurate when compared with ABC and was equal to 82.68%. A J48 Decision Trees [16] model is a data mining algorithm implemented and tested on the 768 samples to detect diabetes. The J48 algorithm was based on a decision tree, which is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data [17]. Rules have been used

to recognize healthy/unhealthy property classes. The results showed that the accuracy of J48 diagnosis was 83.83%.

A Principal Component Analysis (PCA)-Cross Entropy Optimization (CEO)-SVM [18] model, based on SVM and Entropy, was performed and tested on 155 hepatitis patients. SVM was used for classification of samples and the PCA-CE for selecting the features and entropy of effective and important features. The results showed that PCA-CE-SVM model's accuracy was 97.2%. A Multi-Layer Perceptron (MLP) [19] model was a model of ANNs that was performed and tested on 155 samples to detect hepatitis. The MLP model is a three-layer model and its testing and training stages are done by a middle layer. The results showed that the MLP detection accuracy was 91.87%. The Incremental Back Propagation Learning Network (IBPLN) [20] method is a model based on ANN's propagation that has been proposed for the diagnosis of hepatitis of 155 patient samples. The results show that the IBPLN model had a higher detection accuracy compared with KNN, Naive Bayes (NB), SVM and Particle swarm optimization models and its value was 93.34%.

Table 1: Comparison of Models for the Diagnosis of Diseases

Type	Model	Dataset	Instances	Accuracy [%]	Computational Complexity	Ref
Diabetes	PNN	Pima	768	82.37	High	[10]
	FCS-ANTMINER	Pima	768	84.32	High	[11]
	CBR	Data Types	2640	97.67	Medium	[12]
	LDA-ANFIS	Pima	768	84.61	High	[13]
	K-Means-GA-SVM	Pima	768	98.82	Medium	[14]
	MABC	Pima	768	82.68	Low	[15]
	J48	Pima	768	83.83	Low	[16]
Hepatitis	PCA-CE-SVM	Hepatitis	155	97.2	Medium	[17]
	MLP	Hepatitis	155	91.87	Medium	[18]
	IBPLN	Hepatitis	155	93.34	High	[19]
Breast Cancer	Fuzzy-KNN	WBCD	699	99.71	High	[20]
	GONN	WBCD	699	98.24	High	[21]
	CNN	Data Types	3158	82.43	Medium	[22]
	SVM	WBCD	699	97.13	Low	[23]
	KNN	WBCD	699	95.27	Low	[23]
	C4.5	WBCD	699	95.13	Low	[23]
	NB	WBCD	699	95.99	Medium	[23]
Heart	Fuzzy-GA	Cleveland	303	-	High	[24]
	SVM	Data Types	214	85.05	Low	[25]
	MLP	Data Types	214	84.11	Medium	[25]
	RBF	Data Types	214	82.71	Medium	[25]
	BN	Data Types	214	80.37	Low	[25]
	J48	Data Types	214	76.65	Low	[25]
	C4.5	Cleveland	303	86.3	Medium	[26]
	ANN	Cleveland	303	86.6	Medium	[26]

The Fuzzy-KNN [21] model was used on the Wisconsin Breast Cancer Dataset (WBCD) that contains 699 samples tested and implemented to detect breast cancer. The data set was collected from the patients of University of Wisconsin-Madison Hospitals. When the attributes with missing values were removed, a data set with 239 malignant and 444 benign instances was obtained. The fuzzy model was used to infer the rules and KNN was used for data classification. Results showed that the recognition accuracy of Fuzzy-KNN model was 99.71%.

The Genetically Optimized Neural Network (GONN) [22] model, an ANN models, has been proposed for breast cancer detection on 699 samples. Genetic Programming (GP) is used for testing and training of the GONN model. GP is a type of evolutionary algorithm, a subset of machine learning. It initially generates random solutions to solve a problem, and then evolves them based on a Fitness Function (FF). New and improved individuals are produced by applying reproduction, crossover and mutation operators on individuals of the previous generation. Reproduction is an asexual method wherein a selected individual copies itself into the new population. It is effectively the same as one individual surviving into the next generation. Crossover is applied in a GP by simply exchanging sub-trees between two trees, thus forming two new offspring from two parents. Mutation changes a node within a tree or changes its information, thus, affecting only the individual and creating a new solution. Due to crossover and mutation operators, GP techniques do a better job in exploring the search space than other machine learning algorithms. Crossover and mutation operations in GP improve the GONN model. The evaluation was done on the WBCD dataset and results have shown that the detection accuracy of GONN was 98.24%.

The Convolutional Neural Network (CNN) [23] model is inspired from ANN and was implemented and tested over 3,158 data samples. The CNN model is composed of neurons with substantial weight and bias can be learned. Each neuron receives a number of inputs and then it calculates by multiplying the weight of inputs. Finally, using a nonlinear transfer function, it shows the results. The results showed that the diagnostic accuracy was 82.43%. A semi-supervised deep convolutional neural network [23] was developed for breast cancer diagnosis, which used large amount of unlabeled data to improve the accuracy. SVM, KNN, C4.5 and NB models have been implemented and tested on 699 samples for breast cancer detection [24]. The Models' Assessment was conducted in a WEKA environment. The SVM model had a higher accuracy compared to other models.

The Fuzzy-GA [25] hybrid model was implemented and tested on the Cleveland dataset with 303 patients and 75 features for diagnosing heart disease. The GA model was used for FS and the Fuzzy model for deriving laws. In GA for FS, the gene amounts in the chromosomes were selected as 0 and 1. The results showed that the Fuzzy-GA model had high accuracy in making laws. SVM, MLP, RBF, BN and J48 models were conducted and tested on 214 samples of patients with 19 features for the heart disease diagnosis [26]. The results showed that the SVM model had a higher accuracy compared to other models such as the RBF MLP model. The decision support system model [27] based on fuzzy logic was proposed to detect heart disease in 303 samples. The decision support system is a hybrid of C4.5 and ANN. The results show that the diagnosis accuracy in the training phase in both models were 86.3% and 86.6%, respectively. Table 1 summarizes the comparison of the proposed models for the detection of diseases.

3. THE MODIFIED MODEL

A disease can be diagnosed by examining the patient's and other obtained samples. However, for large numbers of patients with multiple tests per patient, an automated decision making system is a valuable tool in clinical practice. This can be achieved by feature selection (FS) methods. FS is the process of extracting the relevant and most informative data from the feature space, so that the feature set is more appropriate for classification. Some features are improper and have no ability to increase the discriminative power of the classifier. Some features are relevant and highly correlated to a specific classification. The evaluation has been done on the Pima Diabetes [28] dataset including 8 characters and 768 samples, the Hepatitis [29] dataset containing 19 characters and 155 samples, the Breast Cancer [30] dataset containing 9 features and 286 samples, and the Heart [31] dataset including 13 features and 180 samples. Each dataset consisted of a set of numeric or categorical attributes (h(1),h(2),h(3),…,h(m),h(m+1)), where m shows predictive attributes and h(m+1) is a class of disease, namely healthy or sick. In Fig. 1, the flowchart of the modified model is shown.

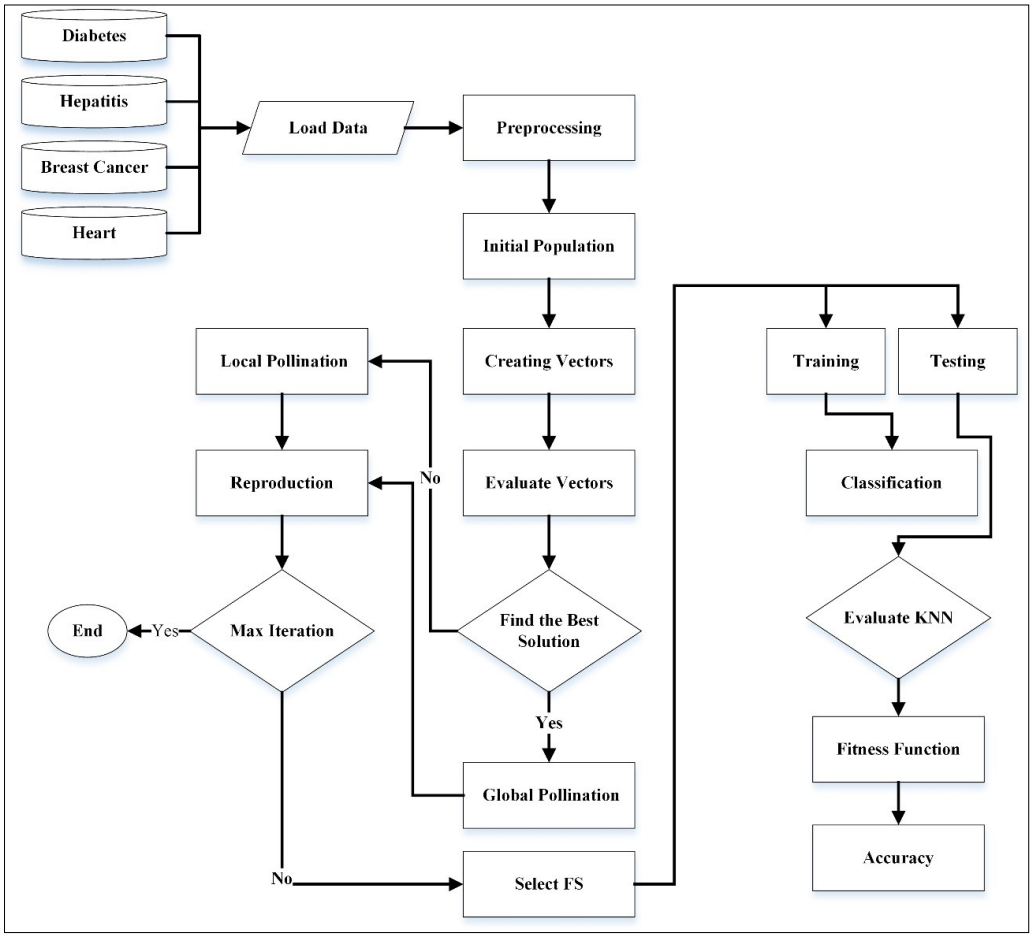

Fig. 1: Flowchart of the modified model.

In the modified model, after importing the raw dataset of disease, was normalized to remove outliers and repetitive data, and then pre-processing operations were done. The initial population was formed as either 0 or 1. The index of the vectors with a value of 1 reflects the selection of the desired feature. Finally, using global pollination operators

based on Eqn. (1) and local pollination based on Eqn. (2) operations, updating and the position changing operations were done on the values [7].

$$x_i^{t+1} = x_i^t + \gamma L(g^* - x_k^t) \tag{1}$$

$$x_i^{t+1} = x_i^t + \varepsilon(x_j^t - x_k^t) \tag{2}$$

In Eqn. (1), x_i^t it is the i^{th} pollen or solution vector x_i at iteration t, and x_i^{t+1} is a candidate solution for iteration t + 1. G is the current best solution, where $g^* = g1, g2, \ldots, gD$. The variable L is the strength of pollination, basically a D-dimensional step size, where $L = l1, l2, \ldots, lD$, and γ is the scaling factor to control the step size. Local Pollination will cause the diversity of weight ranges and poor neighbors run for the classification process. These operators make the weight ranges have closely adjacent properties for KNN classification and they increase the classification accuracy. Since the FS problem is meant to select a specific feature or not, the solution is represented as a binary vector, where 1 indicates a feature will be selected to compose the new dataset and 0 otherwise. A Sigmoid function is used to build this binary vector by the Eqn. (3).

$$S(x_i^j(t)) = \frac{1}{1 + e^{-x_i^j(t)}} \tag{3}$$

Firstly, for each element x_i of the solution vector x_i, a probability $S(x_i^j(t))$ is calculated using the sigmoid function defined by Eqn. (3). Then, using the probability vector $S(x_i^j(t))$, each element in the solution vector $x_i^j(t)$ will either have the value 1 or 0 by applying the condition represented by Eqn. (4). Thus, Eqns. (1) and (3) will be replaced by Eqn. (4).

$$x_i^j(t) = \begin{cases} 1 & if \quad S(x_i^j(t)) > \sigma \\ 0 & otherwise \end{cases} \tag{4}$$

In which $x_i^j(t)$ represents the new pollen (solution) i with the j^{th} feature vector, where $i = 1,2, \ldots, m$ and $j = 1,2, \ldots, d$ at the iteration t and σ ~ U (0, 1). In the modified model, the FF vectors are calculated based on the amount of features that are 1. The vector with higher value of FF is selected, and the vectors that are not optimized undergo local pollination operations to create a new generation of population. Selected features are imported into the KNN model. Training and testing processes are performed in the KNN model. The evaluation is done based on the accuracy of the testing process. The final recognition accuracy is displayed in the output model. In the KNN model, a non-classified sample may easily be found by comparing it with the most similar samples in the training set. Thus, it is necessary to specify criteria for determining the distance between the samples. If we have a feature vector of $< a_1(x), a_2(x), \ldots, a_n(x) >$, to obtain the distance between two features of x_i and x_j according to Eqn. (5) we will use the Euclidean distance [8].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \tag{5}$$

KNN is a model for classifying test samples based on k closest training examples in feature space. The test sample is assigned to the class occurring most often amongst its k

nearest neighbors. Usually, the Euclidean distance is used to measure the closeness of the samples. The KNN model is also a supervised learning-based classification algorithm. In classification, the KNN model is a type of instance-based learning. It classifies objects based on the closest training examples in the feature space. The classification function is only approximated locally and all computation is deferred until classification. The KNN model is amongst the simplest of all machine learning algorithms. The object is classified by a majority vote of its neighbors with the object being assigned to the class most common among its k nearest neighbor (k is positive integer).

In FPA, each feature subset can be seen as a position of pollen. Each subset may contain N features, where N is the number of features in the original set. The fewer the number of features in the solution and the higher the classification accuracy, the better is the solution. Each solution is evaluated according to the modified FF, which depends on two objectives: the solution's accuracy obtained by the KNN classifier and the number of selected features in the solution. The FF used in the modified method is designed to have a balance between the number of selected features in each solution and the classification accuracy obtained by using these selected features, Eqn. (6) represents the FF to evaluate solutions.

$$FF = \mu.acc + \rho\frac{|S|}{|N|} \tag{6}$$

In Eqn. (6), acc expresses the classification error rate of a given classifier KNN is used here. |S| is the cardinality of the selected subset and |N| is the total number of features in the dataset, μ and ρ are two parameters corresponding to the importance of classification quality and subset length. In Eqn. (6), $\mu \in [0,1]$ and $\rho = (1-\mu)$.

The modified model's results should be analyzed in order to determine its value at the evaluation stage, and its methodology identified in its wake. These criteria can be calculated for educational data collection in the learning stage and also for the collection of trial records in evaluation stage. There are various criteria such as *Precision*, *Recall*, *F-Measure* and *Accuracy* for the evaluation and we use accuracy criteria in order to examine the modified model [32, 33].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$F - Measure = \frac{2 * \text{Pr}ecision * \text{Re}call}{(\text{Pr}ecision + \text{Re}call)} \tag{9}$$

$$AUC = \left(\left(\frac{TP}{TP+FN}\right) + \left(\frac{TN}{TN+FP}\right)\right)/2 \tag{10}$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{11}$$

$$ErrorRate = \frac{(FP+FN)}{(TP+TN+FP+FN)} = 1 - Accuracy \tag{12}$$

where the *TN* parameter indicates the number of records with a negative category and that the classification algorithm correctly detected it as a negative one. *TP* represents the number of records that their category is positive and classification algorithm correctly detected it as a positive one. *FP* represents the number of records that their actual category is negative and that the classification algorithm mistakenly detected them as positive. *FN* represents the number of records that their actual category is positive and that the classification algorithm mistakenly detected them as negative.

4. EVALUATION AND RESULTS

In this section, the evaluation and results of the modified model is done on Diabetes, Hepatitis, Breast-Cancer, and Heart disease in the MATLAB 2016 environment. The experiments are tested on an Intel machine Core i7 CPU 2.60 GHz and 6 GB RAM and the values of initial population parameters and the number of iterations are considered as 50 and 100. The values of this parameter are effective in the diagnosis and accuracy of data. To study the impact of the standard parameters of the FPA, a set of extensive experiments has been performed using different values of the main parameters in the algorithm, (i.e., number of initial population and the number of iterations). In Figure (2), an overview of program implementation is shown.

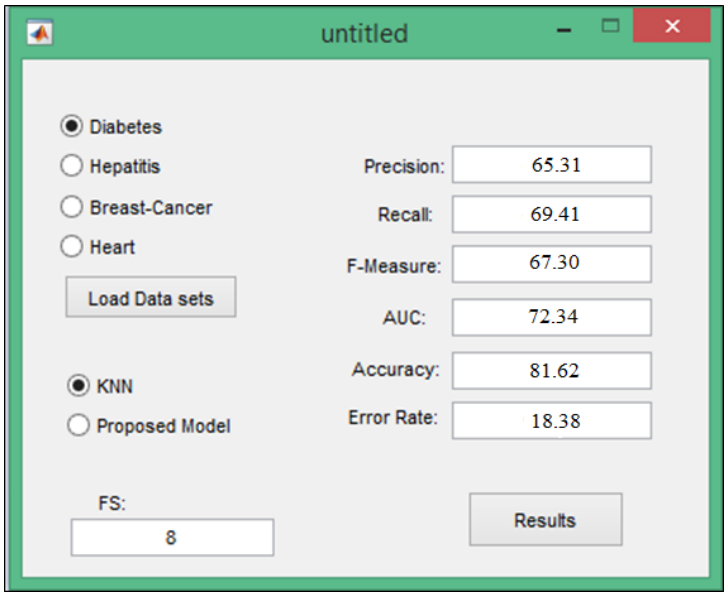

Fig. 2: Overview of Implemented Program in MATLAB 2016a Environment.

In Table (2), the results of the models on different datasets are shown. As shown in Table (2), each criterion has a different value based on the data within the dataset. The modified model was much more accurate and had a lower Error Rate than KNN models. The accuracy of the modified model for Diabetes, Hepatitis, Breast-Cancer, and Heart disease datasets was 86.34%, 85.13%, 82.39%, and 83.04%, respectively.

In Figure (3), a comparison chart of the models based on the Error Rate criteria is shown. In Figure (3), it is clear that the modified model has less error than the KNN model.

Table 2: Results of Models on Different Datasets

Datasets	Models	Precision	Recall	F-Measure	AUC	Accuracy	Error Rate
Diabetes	KNN	65.31	69.41	67.30	72.34	81.62	18.38
	Modified Model	71.34	75.49	73.36	79.24	86.34	13.66
Hepatitis	KNN	66.38	67.48	66.93	71.89	79.20	20.80
	Modified Model	73.61	78.01	75.75	80.64	85.13	14.87
Breast-Cancer	KNN	66.21	68.40	69.27	70.05	73.48	26.52
	Modified Model	69.15	74.68	71.81	79.60	82.39	17.61
Heart	KNN	70.14	71.45	70.79	72.85	76.46	23.54
	Modified Model	73.16	76.11	74.61	79.50	83.04	16.96

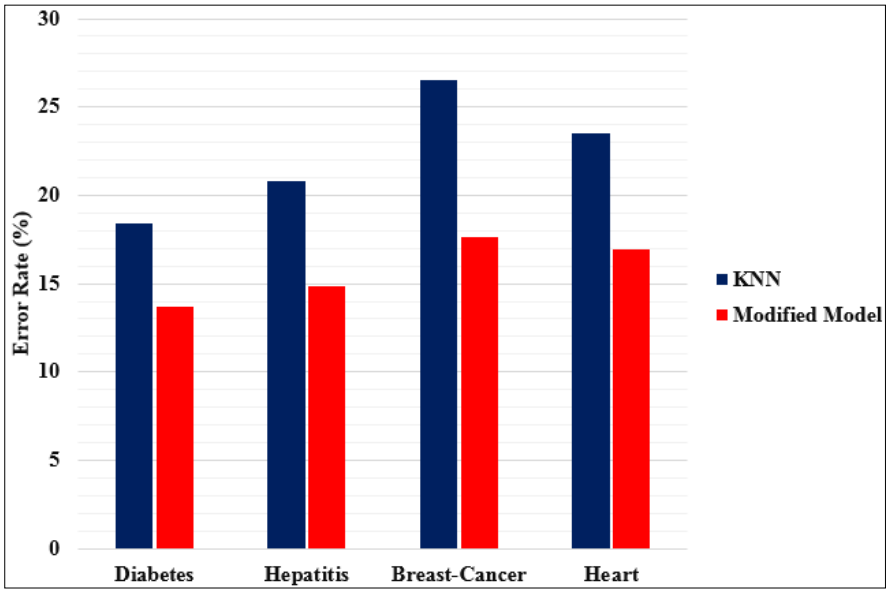

Fig. 3: Comparison Chart of models based on the Error Rate.

In Table 3, the results of the modified model are shown based on different iterations on the datasets. By increasing iterations, the accuracy value has been increased and the error value has been decreased. The accuracy of the modified model with 500 iterations for Diabetes, Hepatitis, Breast-Cancer, and Heart disease datasets was 95.76%, 94.79%, 89.90%, and 90.03%, respectively.

In Table 4, the comparison of models based on FS with 500 iterations is shown. In Table 4, a comparison of models is shown, based on the FS. As shown in Table 4, when the number of features is low, then the accuracy is better because fewer features need to be evaluated, but the features that have a greater impact on diagnosis and fewer errors are selected.

Table 3: The Results of the Modified Model based on Iterations on Different Datasets

Datasets	Iterations	Precision	Recall	F-Measure	Accuracy	Error Rate
Diabetes	100	71.34	75.49	73.36	86.34	13.66
	150	72.38	76.48	74.34	89.82	10.18
	200	74.65	76.98	75.80	92.37	7.63
	300	75.63	77.15	76.38	93.47	6.53
	500	50.25	81.49	80.87	95.79	4.21
Hepatitis	100	73.61	78.01	75.75	85.13	14.87
	150	75.38	77.52	76.44	88.34	11.66
	200	78.95	79.14	79.04	90.80	9.20
	300	80.32	82.31	81.30	91.46	8.54
	500	83.79	85.19	84.48	94.79	5.21
Breast-Cancer	100	69.15	74.68	71.81	82.39	17.61
	150	70.16	71.62	70.88	86.56	13.44
	200	71.49	73.18	72.33	88.17	11.83
	300	71.49	72.08	72.01	89.21	10.79
	500	72.50	74.10	73.29	89.90	10.10
Heart	100	73.16	76.11	74.61	83.04	16.96
	150	74.53	75.20	74.86	85.16	14.84
	200	82.47	85.14	83.78	88.94	11.06
	300	85.95	87.46	86.70	89.63	10.37
	500	86.41	85.21	85.81	90.03	9.97

Table 4: Evaluation and Results of the Modified Model on Different Datasets based on the FS

Datasets	FS	Precision	Recall	F-Measure	AUC	Accuracy	Error Rate
Diabetes	5	73.94	75.81	74.86	85.16	**96.35**	3.47
	6	73.00	74.81	73.89	83.20	93.18	6.82
	7	72.14	73.22	72.68	82.07	92.16	7.84
	8	71.34	75.49	73.36	79.24	89.34	10.66
Hepatitis	10	77.00	79.01	77.99	86.00	**94.05**	5.95
	12	75.13	76.31	75.72	84.07	91.02	8.98
	16	74.82	75.89	75.35	72.16	90.64	9.36
	19	73.61	78.01	75.75	80.64	90.13	9.87
Breast-Cancer	5	74.20	76.21	75.19	83.09	**91.16**	8.84
	6	72.61	73.91	73.25	81.44	89.10	10.90
	8	70.50	72.46	71.47	80.17	86.88	13.12
	9	69.15	74.68	71.81	79.60	86.39	13.61
Heart	9	78.35	79.05	78.70	86.12	**94.34**	5.66
	11	75.14	77.68	76.39	83.17	92.19	7.81
	12	74.91	75.41	75.16	80.16	89.10	10.90
	13	73.16	76.11	74.61	79.50	87.04	12.96

In Table (4), the accuracy for Diabetes with FS=5 and without FS is 96.35% and 89.34%, respectively. The accuracy for Hepatitis with FS=10 and without FS is 94.05% and 90.13%, respectively. The accuracy for Breast Cancer with FS=5 and without FS is 91.16% and 86.39%, respectively. As well, in Table (4), the accuracy for Heart with FS=9 and without FS is 94.34% and 87.04%, respectively. In Figure (4), a comparison chart of the modified model based on FS is shown. In Figure (4) it is clear that if the number of features is low, then the Error Rate will also be lower, because FS is directly related to the Error Rate and it is effective in determining accuracy error.

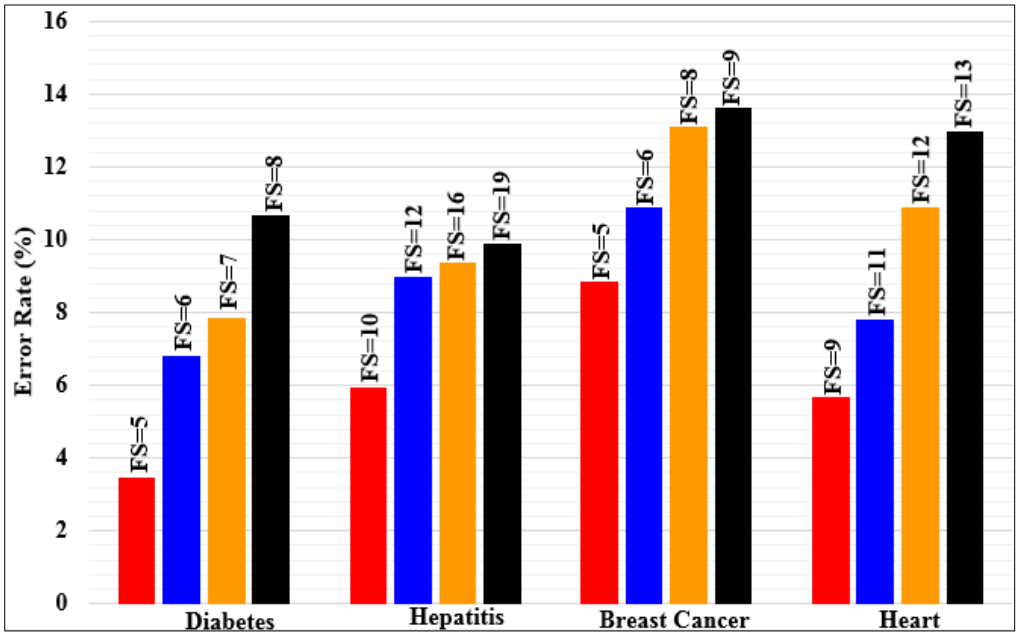

Fig. 4: Comparison Chart of the Modified Model based on FS.

5. CONCLUSION AND FUTURE WORK

Some diseases are diagnosed early and appropriate treatment for them would be considered. Unfortunately, in the world 70 to 90 percent of diseases are not diagnosed in time and patients are sick (advanced illness) before visiting the doctor, when there would no longer be an appropriate treatment for them. In this paper, by improving the KNN and using FPA, a model for predicting diseases was developed. In the modified model, the FS method was used to increase the efficiency and accuracy of the diagnosis system. FS can identify the important features that can help doctors to diagnose the disease in the earliest instance. Results on the datasets of Diabetes, Hepatitis, Breast Cancer and Heart disease showed that the modified model has better recognition accuracy than a KNN model. If the number of features is less, the processing will be better and also the key features will be extracted. As future work, we intend to construct a computational model in which, instead of combining feature subsets, we will combine the mathematical formulas that define the feature selection metrics with the goal to achieve a more impartial formula.

REFERENCES

[1] Begum S, Ahmed MU, Funk P, Xiong N, Folke M. (2011) Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(4):421-434.

[2] Srikanth P, Deverapalli D. (2016) A Critical Study of Classification Algorithms Using Diabetes Diagnosis. in 2016 IEEE 6th International Conference on Advanced Computing (IACC).
[3] Neshat M, Sargolzaei M, Nadjaran Toosi A, Masoumi A. (2012) Hepatitis Disease Diagnosis Using Hybrid Case Based Reasoning and Particle Swarm Optimization. ISRN Artificial Intelligence. 2012:6.
[4] Bhatia S, Prakash P, Pillai GN. (2009) Svm Based Decision Support System for Heart Disease Classification with Integer-coded Genetic Algorithm to Select Critical Features.
[5] Arya C, Tiwari R. (2016) Expert system for breast cancer diagnosis: A survey. in 2016 International Conference on Computer Communication and Informatics (ICCCI).
[6] Bazazeh D, Shubair R. (2016) Comparative study of machine learning algorithms for breast cancer detection and diagnosis. in 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA).
[7] Yang X-S. (2012) Flower Pollination Algorithm for Global Optimization. in Unconventional Computation and Natural Computation. Berlin, Heidelberg: Springer Berlin Heidelberg.
[8] Martin B. (1995) Instance-based learning: nearest neighbour with generalisation, in Computer Science Working Papers. University of Waikato, Department of Computer Science.
[9] Duda RO HP, Stork DG. (2007) Pattern Classification. vol. 24. John Wiley & Sons.
[10] Temurtas H, Yumusak N, Temurtas F. (2009) A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4):8610-8615.
[11] Ganji MF, Abadeh MS. (2011) A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 38(12):14650-14659.
[12] El-Sappagh S, Elmogy M, Riad AM. (2015) A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. Artificial Intelligence in Medicine, 65(3):179-208.
[13] Dogantekin E, Dogantekin A, Avci D, Avci L. (2010) An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS. Digital Signal Processing, 20(4):1248-1255.
[14] Santhanam T, Padmavathi MS. (2015) Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. Procedia Computer Science, 47:76-83.
[15] Beloufa F, Chikh MA. (2013) Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer Methods and Programs in Biomedicine, 112(1):92-103.
[16] Hayashi Y, Yukita S. (2016) Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked, 2:92-104.
[17] Gashti MZ. (2017) A novel hybrid support vector machine with decision tree for data classification. International Journal of Advanced and Applied Sciences, 4(9):138-143.
[18] Elaboudi N, Benhlima L. (2016) A New Approach Based on PCA and CE-SVM for Hepatitis Diagnosis. in Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015. Cham: Springer International Publishing.
[19] Bascil MS, Temurtas F. (2011) A study on hepatitis disease diagnosis using multilayer neural network with levenberg marquardt training algorithm. J Med Syst, 35(3):433-436.
[20] Mitra M, Samanta RK. (2015) Hepatitis Disease Diagnosis Using Multiple Imputation and Neural Network with Rough Set Feature Reduction. in Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Cham: Springer International Publishing.
[21] Onan A. (2015) A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications, 42(20):6844-6852.
[22] Bhardwaj A, Tiwari A. (2015) Breast cancer diagnosis using Genetically Optimized Neural Network model. Expert Systems with Applications, 42(10):4611-4620.

[23] Sun W, Tseng TB, Zhang J, and Qian W. (2017) Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Comput Med Imaging Graph, 57:4-9.
[24] Asri H, Mousannif H, Moatassime HA, Noel T. (2016) Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83:1064-1069.
[25] Ephzibah EP, Sundarapandian V. (2012) A Fuzzy Rule Based Expert System for Effective Heart Disease Diagnosis. in Advances in Computer Science and Information Technology. Computer Science and Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg.
[26] Ghumbre SU, Ghatol AA. (2012) Heart Disease Diagnosis Using Machine Learning Algorithm. in Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. Berlin, Heidelberg: Springer Berlin Heidelberg.
[27] Lahsasna A, Ainon RN, Zainuddin R, Bulgiba AM. (2012) A Transparent Fuzzy Rule-Based Clinical Decision Support System for Heart Disease Diagnosis. in Knowledge Technology. Berlin, Heidelberg: Springer Berlin Heidelberg.
[28] Pima Indians Diabetes Dataset.[https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes]
[29] Hepatitis Dataset.[https://archive.ics.uci.edu/ml/datasets/Hepatitis]
[30] Breast Cancer Dataset.[http://www.sgi.com/tech/mlc/db/breast-cancer.all]
[31] Heart Diseases Dataset.[http://www.sgi.com/tech/mlc/db/heart.data]
[32] Farjamnia G, Gashti MZ, Barangi H, and Gasimov YS. (2017) The Study of Support Vector Machine to Classify the Medical Data. IJCSNS International Journal of Computer Science and Network Security, 17(12):145-150.
[33] Ryszard SM, Ivan B, Avan B. (1998) Machine Learning and Data Mining; Methods and Applications. John Wiley \& Sons, Inc. 456.