Neural Processing Letters (2023) 55:153–169
https://doi.org/10.1007/s11063-021-10491-0

Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm

Xiaohua Li[1] · **Jusheng Zhang**[2,3] · **Fatemeh Safara**[4]

Accepted: 9 March 2021 / Published online: 27 March 2021

Abstract
Artificial intelligence is a future and valuable tool for early disease recognition and support in patient condition monitoring. It can increase the reliability of the cure and decision making by developing useful systems and algorithms. Healthcare workers, especially nurses and physicians, are overworked due to a massive and unexpected increase in the number of patients during the coronavirus pandemic. In such situations, artificial intelligence techniques could be used to diagnose a patient with life-threatening illnesses. In particular, diseases that increase the risk of hospitalization and death in coronavirus patients, such as high blood pressure, heart disease and diabetes, should be diagnosed at an early stage. This article focuses on diagnosing a diabetic patient through data mining techniques. If we are able to diagnose diabetes in the early stages of the disease, we can force patients to stay home and care for their health, so the risk of being infected with the coronavirus would be reduced. The proposed method has three steps: preprocessing, feature selection and classification. Several combinations of Harmony search algorithm, genetic algorithm, and particle swarm optimization algorithm are examined with K-means for feature selection. The combinations have not examined before for diabetes diagnosis applications. K-nearest neighbor is used for classification of the diabetes dataset. Sensitivity, specificity, and accuracy have been measured to evaluate the results. The results achieved indicate that the proposed method with an accuracy of 91.65% outperformed the results of the earlier methods examined in this article.

Keywords Diabetes diagnosis application · Genetic algorithm · Particle swarm optimization · Harmony search algorithm · K-means · Artificial intelligence · Coronavirus disease pandemic

1 Introduction

Accurate and early diagnosis of a disease plays a vital role in the provision of treatment. This may also have an impact on social health and life satisfaction. A great deal of research has been carried out to develop a computer-aided diagnostic system based on artificial intelligence

✉ Jusheng Zhang
18675088889@163.com

Extended author information available on the last page of the article

techniques and decision support systems [1]. Because some diseases exhibit common symptoms, the diagnostic process is sometimes complex and optimization techniques are required to overcome the problem. Metaheuristic and optimization algorithms are widely used techniques for offering intelligent diagnostic systems to enhance classification accuracy [2, 3]. Most metaheuristic algorithms receive their inspiration from nature to understand complex relationships from basic and simple conditions. Nature is a perfect example of optimization. Each metaheuristic algorithm creates an initial population of applied solutions and then moves repeatedly from generation to generation in the direction of the best solution [4].

People with diabetes are prone to have serious complications such as cardiovascular diseases, eye conditions, and coronavirus disease (COVID-19). With a high prevalence of COVID-19, people with diabetes are exposed to an increased risk of being infected with the disease [5]. Given a large number of people with diabetes in today's societies, and the fact that some people are unaware that they have diabetes, to diagnose diabetes is of particular importance [6, 7]. It may reduce the number of diabetic patients suffering from COVID-19. It is a metabolic condition that is one of the most common chronic diseases in today's modern society. In people with diabetes, the pancreas' ability to produce Insulin is weakened or stopped, or the body does not absorb the insulin. Therefore, the body cannot perform its metabolism correctly. Insulin has the crucial role in reducing the levels of blood sugar through different mechanisms. A prolonged increase in blood sugar in the body can lead to side effects such as blindness, kidney disease, cardiovascular disease, and neurosis.

Regarding the high prevalence of this diabetes, it was implementing a method that can determine the correct diagnosis of whether or not to have diabetes can be an essential step in diagnosis and controlling the disease. It can lead to the detection, prevention and treatment of COVID-19 [8], because individuals with diabetes are more likely to contract COVID-19 than the general population [9]. That's why medical professionals need a reliable method or predictive medical system to diagnose diabetes. It is hoped that by monitoring the effects of diabetes, the cost of treatment will be reduced. Data mining and evolutionary algorithms are some of these useful tools. The aim of this study is to diagnose diabetes using data mining and intelligent metaheuristic algorithms. The main contributions of this paper are as follows:

- K-means clustering algorithm is used to improve the accuracy of feature selection.
- Metaheuristic Harmony search algorithm is proposed and examined for feature selection.
- K-Nearest Neighbor (KNN) is used for classification, and different combinations of KNN and Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Harmony search algorithm are examined for diabetes disease dataset classification.

The remainder of that article is organized as follows: recent researches reported on diabetes diagnosis are represented in Sect. 2. Details of the proposed method and evaluation parameters are explained in Sect. 3. Section 4 is dedicated to a description of the dataset used and results gained through the proposed method. The results are analyzed and discussed in the same section as well. The conclusion is provided in Sect. 5.

2 Related Works

An significant part of the researches is on diseases caused by diabetes [10]. Sambyal et al. [11] conducted a study on using deep learning to predict eye, kidney, and cardiovascular diseases. They evaluated their results using statistical analysis. Data mining techniques

are used for appropriate screening, evaluating, prediction, and following up of the current patients and probable future patients with different diseases [8, 12, 13]. Guo et al. [14] found that patients with COVID-19 and diabetes, and no other diseases had a high risk of serious lung diseases such as pneumonia and conditions related to enzymes and blood particles. Their results confirmed the idea that diabetics are at high risk for a fast progression of COVID-19. Therefore, more intensive attention should be paid to provide treatments for both current diabetic people and the people who are prune to be diabetics. Several research articles have been published on diagnosis diabetes on different diabetes dataset. One of the common and widely used dataset is the PIMA Indian Diabetes dataset available from the UCI repository. Some of the publications presented a comparative study on the impact of the different data mining techniques, and other publications are presented to propose an improvement on current methods and results in a new method for diabetes prediction [15, 16].

Several studies were reported on the different dataset as PIMA Indian diabetes dataset. One of them is [17], where the use of the convolutional neural network and the long-term memory for diagnosing diabetes is proposed. The experiments were based on the patient's heart rate extracted from their electrocardiogram (ECG), which results in the accuracy of 93.6%. The other one is the research presented by Mirza et al. [1]. They developed a prediction model for diabetes prediction employing SMOTE, and Decision tree (DT) classification algorithms. SMOTE is used due to its capability of managing imbalanced data. They combined DT and SMOTE intending to improve an accuracy of diabetic prediction by eliminating class imbalance. With the hybrid method, the classification accuracy of 94.70% is achieved on the dataset collected by the authors. In [18] several algorithms are examined on the PIMA Indian dataset and a localized dataset. Principle component analysis (PCA) and PSO is also used in different combinations with classification algorithms. The best results of 79.56% by PCA-LR and 92.43% by PSO-Naive Bayes were achieved on the PIMA Indian and localized datasets. The PSO is also employed by [19], to improve ANN accuracy for diabetes detection. They successfully tried to control the saturation rate of PSO activation function.

A comparative review is performed by Ganesh et al. [20] on diabetes classification. The review paper reported the research studies conducted on PIMA Indian diabetes dataset using artificial neural network (ANN) and support vector machine (SVM). The advantages and disadvantages of the researches reviewed were presented and their accuracies are compared. A review article is published by Gujral et al. [21] that compared the methods used to diagnose diabetes on the PIMA Indian dataset. The methods and algorithms considered include fuzzy logic, SVM, GA, ANN, and PCA classification algorithms. Similarly, in [22], the diagnosis of diabetes is based on the PIMA Indian dataset, which provides an overview of the methods, weaknesses and strengths along with their results. SVM, ANN, Naïve Bayesian, J48 Decision tree, Bagging method, and combined method of GA with SVM.

In [23], the authors compare machine learning classifiers, including J48 Decision Tree, Random Forest, KNN, and SVM to classify patients with diabetes mellitus. The classifiers are examined on the machine learning repository. Two evaluation processes have been conducted, one on the dataset with a noisy dataset which is without any preprocessing and the other one on clean data which is after preprocessing. Two versions of ANN, including stochastic and pruned stochastic are proposed in [24], and tested on 60 datasets, one of the datasets is about diabetes, and their performances are evaluated.

In [25], a method is proposed for predicting diabetes using Adaboost, which also performs an evaluation analysis. The use of the Adaboost is based on decision tree C4.5

that has been used in the training stage. The results indicate the efficiency of their proposed method combining Adaboost and C4.5 over the Boosting and Begging methods.

In [26], a fuzzy classification method is proposed using a bee colony optimization (BCO) algorithm for diabetes. The modified BCO algorithm was used for the creation and optimization of the membership functions and rules derived from the data. Classification accuracy of 84.21% is achieved on the PIMA Indian dataset.

Ant colony optimization is also employed by Singh et al. [27] for PIMA Indian diabetes dataset classification combined with a fuzzy rule-based system. Although different combinations of algorithms were examined in the paper, the best accuracy of 87.7% was achieved by ANT-FDCSM. ANT-FDCSM is a rule-based metaheuristic algorithm which was derived from the ant colony optimization algorithm. An optimal split point selection algorithm is used to improve the algorithm, and tenfold cross validation was used to evaluate the training and test results. Intuitionistic Fuzzy SVM is proposed by Laxmi et al. [28] and examined on different dataset, including the diabetes dataset, and its applicability is shown.

Hassan et al. [29] examined a self-organizing map (SOM) optimization algorithm with four metaheuristic algorithms, including PSO, newton-based SOMPSO, SOMHSA (SOM with Harmony search algorithm), and SOMSwram. The best accuracy of diagnosis of diabetic patients of 80% is achieved on PIMA Indian diabetes dataset. The four algorithms are also examined on Wisconsin and newThyroid dataset, and better accuracies than those on the PIMA Indian dataset were obtained. For example, for the new Thyroid dataset, accuracy of 91% through newton-based SOM, and Wisconsin dataset, accuracy of 97% was gained through SOMHSA.

In [30], early diagnosis of type II diabetes has been done using multiple classification systems, called multiple factors weighted combination (MFWC), to develop the accuracy of diagnosis for complex type II diabetes, dynamic weighing schemata, known as the weighted combination of multiple criteria, is presented to combine classification in decision making.

In [31], a powerful intelligent diagnostic system for diabetics is presented from the PIMA Indian dataset based on a hybrid algorithm called Logistic Adaptive Network-based Fuzzy Inference System (LANFIS). The accuracy of the proposed methodology is relatively insufficient, estimated at 88.03% and its computer complexity is high. Kannadasan et al. [32] proposed a Deep Neural Network (DNN) framework for diabetes data classification. In the paper, feature selection is performed by stacked auto-encoders. The DNN framework is applied to the PIMA Indian diabetes and obtained an accuracy of 86.26% of classification.

Choubey et al. [33] classified the diabetes dataset through variety of algorithms, including KNN, Adaboost, and ANN with Radial Basis Function (RBFN). Then, to improve the precision of the classification algorithm above, they looked at the selection of features through linear discriminant analysis (LDA) and PCA. Therefore, each of the classification algorithms is performed after one of the PCA or LDA. Among all classification combinations examined, PCA and CVR were the most accurate for the two datasets. Several data mining techniques are examined by Al-Zebrai et al. [34], including SVM, Coarse Gaussian SVM, LDA, Decision Trees (DT), KNN, Logistic Regressions (LR), and ensemble learners. In the experiment, LR with the 77.9% accuracy was the best classifier, and Coarse Gaussian SVM with the 65.5% accuracy was the worst.

In the articles reviewed, the common dataset used to assess proposed methods for the diagnosis of diabetes is the PIMA dataset on diabetes in India. It makes the evaluation

process of newly proposed algorithms easier. Although many papers are published on diabetes diagnosis, the accuracy of diagnosis has still needed to be improved.

3 Materials and Methods

In this section, first, classification, clustering, and metaheuristic algorithms used in this paper are explained briefly. Then, the proposed method that is designed based on the algorithms is described. Evaluation parameters are provided at the end of this section as well. All mathematical symbols used in the equations of this section are presented in Table 5 in "Appendix A".

3.1 Metaheuristic and Data Mining Algorithms

In this article, Harmony, GA and PSO as metaheuristic algorithms are used in conjunction with K-means for feature selection in order to improve the accuracy of KNN by providing some discriminatory features. A brief explanation of each algorithm is provided first, followed by an explanation of the proposed method.

Genetic Algorithm Genetic algorithm is a metaheuristic search optimization algorithm working on Darwin's survival theory [35]. GA inspires from the process of natural choice that the best individuals are chosen to be used for producing offspring of the subsequent generation. GA works as explained in the following:

1. GA starts generating an initial random population.
2. The algorithm generates a series of new populations. The current population is employed to generate the subsequent population at each iteration. Following steps, a to f, should be followed to generate the new population:
3. The algorithm ends when one of the terminating conditions, time limit or fitness limits, are satisfied.

    a. The algorithm scores each member of the current population by calculating corresponding fitness values, which are named as the raw fitness scores.
    b. The algorithm scales the raw fitness scores to normalize the scores into a more practical range of values. These scaled values are called expectation values.
    c. The algorithm selects members that are called parents, based on their expectation.
    d. The individuals with lower fitness in the current population are selected as elite, and passed to the next population.
    e. The algorithm generates children from their parents. Children are generated either by applying random changes to a single parent, which is called a mutation, or by merging the vector entries of a pair of parents, which is called crossover.
    f. The algorithm substitutes the current population with their children to produce the next generation.

Harmony Harmony is an optimization algorithm that is introduced in 2001. The algorithm was inspired by the work of musicians to enhance the instrument's performance. Harmony optimization algorithm works as GA from the selecting best individuals' point of view. It works as GA in generating the next Harmony generation based on the current population as well. The goal of Harmony search algorithm is finding the best response from a set of responses. It is

similar to seeking the best solution to transform the quantitative approach into a qualitative approach and improve end results. The major limitation of the Harmony search algorithm is that it does not work well with high-dimensional data [36]. However, the dataset used in this article is not high-dimensional, and the Harmony search algorithm worked well. In this article, we used a Harmony search algorithm to find the best features of detecting diabetes among all features.

With the Harmony search algorithm, a parameter of accepting rate r_{accept} is defined that its values are from the range of [0, 1]. The low varying r_{accept} can result in slow convergence, and high varying r_{accept}, can cause some of the harmonies not be explored, and final results would not to be satisfactory. Therefore, the usual range of r_{accept} is 0.7 to 0.95.

The other component is P_{pitch} which is the pitch adjustment. The P_{pitch} is determined by b_{range}, which is the pitch bandwidth, and r_{pa}, which is the adjusting rate, as presented in Eq. (1).

$$x_{new} = x_{old} + b_{range} * \varepsilon \tag{1}$$

x_{old} is the current pitch of a Harmony algorithm, and x_{new} is the pitch after adjusting. x_{new} is the new solution that is obtained from the previous solution by altering the current pitch. ε is a random number between 0 and 1. The operation of the adjusting pitch is similar to the mutation in GA. By the adjusting r_{pa} we can control the adjustment degree. If the adjusting rate is low and the bandwidth is narrow, the convergence of the Harmony search algorithm would be slow, and vice versa. Therefore, r_{pa} is usually determined from the range of 0.1 to 0.5.

P_{random} is the third component that is used to produce more variations of a solution that can be calculated as formulated in Eqs. (2) and (3):

$$P_{random} = 1 - r_{accept} \tag{2}$$

Thus, the probability of adjusting pitches would be as follows:

$$P_{pitch} = r_{accept} * r_{pa} \tag{3}$$

The advantage of this algorithm is the combination of different solutions using a vector created from the best individuals. Several combinations help the search for the best value found. The time needed for the combination could be a disadvantage of the algorithm.

Particle Swarm Optimization Inspiring by fish shoaling and bird flocking social behavior, Eberhart and Kennedy [37] proposed PSO stochastic optimization algorithm. In the PSO, the particles indicate each component of the folk elements, which defines physical properties such as mass and volume. Following equations, (4) to (7), provide the way PSO works [32].

$$X_i = \left(x_{i1}, x_{i2}, \ldots, x_{iD}\right) \tag{4}$$

$$P_i = \left(p_{i1}, p_{i2}, \ldots, p_{iD}\right) \tag{5}$$

$$V_i = \left(v_{i1}, v_{i2}, \ldots, v_{iD}\right) \tag{6}$$

$$v_{id} = w * v_{id} + c_1 * r_1 * \left(P_{id} - X_{id}\right) + c_2 * r_2 * \left(P_{gd} - X_{id}\right) \tag{7}$$

where the current position of a particle is x_{id}, the p_{best} of the particle is P_{id}, the g_{best} of the group is P_{gd}, the velocity of particle is v_{id}, the interia factor is w, the relative influence of

the cognitive component is c_1, the relative influence of the social component is c_2, and r_1, r_2 are random numbers. r_1, r_2 are employed to keep the population's change spread between [0, 1], equally. The c_1 and c_2 are the self-recognition t constant and the social component coefficient, as shown in Eq. (8).

$$w = w_{max} - \frac{w_{max} - w_{min}}{iter_{max}} \tag{8}$$

where the initial weight is shown by w_{max}, the final weight is shown by w_{min}, the maximum iteration number is shown by $iter_{max}$, and the current iteration number is shown by iter.

K-Nearest Neighbors The following steps are passed by the KNN to classify the instances of a dataset:

1. Determine the number K of the neighbors.
2. Compute the distances between the desired data-point and its K neighbors using the Euclidean distance. The Euclidian distance can be calculated as depicted in Eq. (9):

$$d(p, q) = \sqrt{\left(p^1 - p^2\right) + \left(q^1 + q^2\right)} \tag{9}$$

   where p and q are the data-points, we are going to compute their distance.
3. Select the K nearest neighbors in terms of Euclidean distance calculated.
4. Count the number of data-points of each class from the k neighbors selected in the previous step.
5. Put the new data-points to the category in which the number of the neighbor is maximum.

K-means The K-means algorithm follows the next steps to cluster the records:

1. Set k random points as means.
2. Place each item into a group of items with the closest average and update it to include the new item.
3. Repeat the process to meet the stopping criteria.

Several approaches could be taken to initialize the means. Instead of means of random data points, the means can be calculated from the data points which are at the boundaries of the dataset. The main goal of a clustering method is to find similarities between records to put them in one cluster and to find dissimilarities between two records to put the records in different clusters. Therefore, all the records of a cluster have almost the same properties. This within-cluster similarities could help the feature selection method in looking for a more useful and informative feature from the records of one cluster. In other words, clustering is performed to make clusters ready for GA, PSO, and Harmony search algorithm for finding relationships between records and used these relationships to rank the features and select proper features.

3.2 Proposed Method

Feature selection is an essential stage before classification that affects the results of classification considerably. In this paper, combinations of different metaheuristic algorithms are examined to improve the accuracy of KNN in diabetes diagnosis. Three hybrids are

examined: GA-Kmeans, GA-PSO-Kmeans, and Harmony-Kmeans (HR-Kmeans); (Note: To make reading the name of hybrid algorithms easier, K-means is written as Kmeans). Then, KNN is used for classification.

3.2.1 Feature Selection through Hybrid GA and PSO, HR and K-Means

Feature selection is a widely used technique in the classification applications. The quality of classification result is highly dependent on the selected features. In the feature selection process, noisy and redundant features would be removed while informative features would be preserved [38]. Three primary feature selection methods are the filter method, wrapper method, and embedded method [39]. The filter method statistically studies the inherent characteristics of the dataset and calculates the score for each feature regardless of any classifier outcome. Wrapper method scores features based on their usefulness to improve classification performance. In the embedded method, the search process for the best feature subset is indirectly integrated into classifier construction, such as the decision tree. In this paper, the wrapper method is used.

K-means is used to divide the whole dataset into two clusters. Although the dataset was labelled, at this stage of the proposed approach, labels were not considered. We tried to double-check the discriminatory power of the final selected features. As we examined the K-means clustering algorithm with metaheuristic algorithms for one stage feature selection first, and then we applied the KNN classification algorithm to check, as the second stage, the appropriateness of the features selected.

Three hybrids of metaheuristic algorithms are examined, GA-Kmeans, GA-PSO-Kmeans and HR-Kmeans, as explained in the followings:

GA-Kmeans As the first hybrid of algorithms, we combined GA and K-means clustering. First, clustering is performed, and all of the dataset records are separated into two clusters. Then, GA is used to investigate the relationship between the records in each cluster and to determine the impact of each feature to assign a record to a cluster. The trend of decreasing fitness function throughout 200 iterations is illustrated in Fig. 1.

Then, all the records fed into a KNN to be classified concerning their labels into two classes: people who have diabetes and people who do not have diabetes.

GA-PSO-Kmeans The same procedure was repeated with the combination of GA and PSO for feature selection. GA is a common and widely used metaheuristic algorithm for feature selection in classification applications. However, it lacks a high convergence rate. To increase the convergence rate of GA, we employed PSO, which has a high convergence rate. The main drawback of PSO is getting stuck in local minima. GA and PSO are used together in order to decrease the chance of getting stuck in local minima. Combination of GA and PSO were examined before in the different fields of studies, including medical and engineering applications. However, the combination did not examine for diabetes diagnosis. Besides, we applied GA-PSO in clusters provided by K-means, which is another contribution of this paper.

The trend of decreasing fitness function while using the GA-PSO-Kmeans feature selection is shown in Fig. 2. The least fitness function value is less than that of the GA-Kmeans combination.

HR-Kmeans As the third feature selection examination, we employed Harmony search algorithm to find proper features from each of the clusters provided by the K-means algorithm. Then, KNN is used to divide the dataset records into two classes: people with diabetes and people without diabetes. The decreasing trend of fitness function

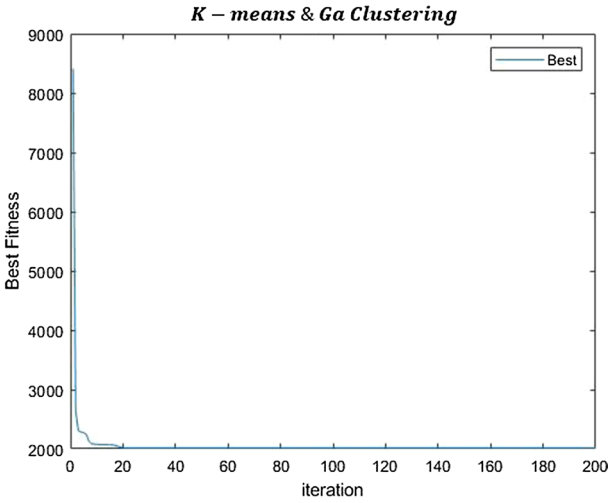

Fig. 1 Decreasing the fitness function by the GA-Kmeans hybrid

values while using Harmony feature selection is shown in Fig. 3. The least fitness function value is less than that of both GA-Kmeans and GA-PSO-Kmeans combinations.

3.2.2 Classification

The features selected in the previous section by three different hybrids of GA, PSO, and Harmony metaheuristic algorithms with K-means clustering, were fed into KNN. KNN was examined before for diabetes diagnosis on the same dataset, PIMA Indian diabetes dataset. However, the results achieved were not convincing. We tried to improve the accuracy of KNN classification through enhanced feature selection. The results provided in the next section approved the appropriateness of the proposed feature selection methods.

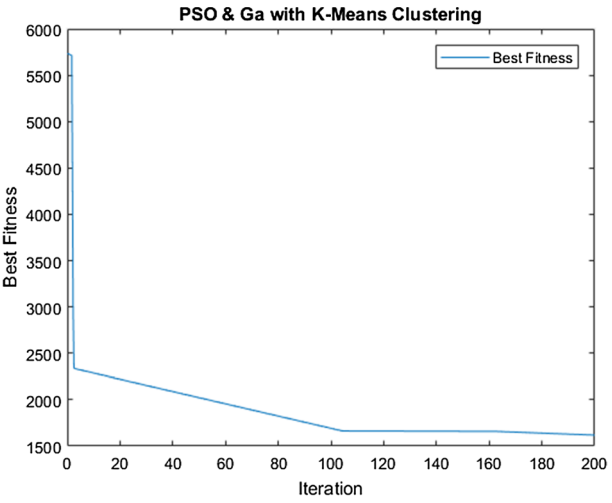

Fig. 2 Decreasing the fitness function by GA-PSO-Kmeans hybrid

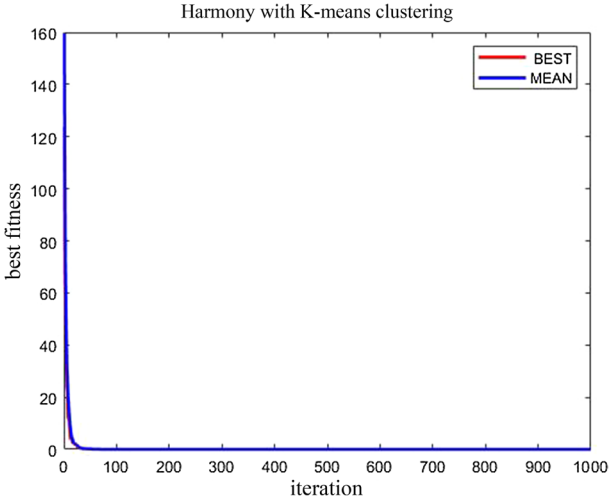

Fig. 3 Decreasing the fitness function by HR-Kmeans hybrid

3.3 Evaluation Parameters

Three widely used parameters, accuracy, sensitivity, and specificity, presented in Eqs. (10) to (12), to evaluate machine learning algorithms are employed in this paper to show the capability of the proposed method to diagnose diabetes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FP} \tag{12}$$

TP is the rate of records that have diabetes and classified as diabetic people correctly. TN is the rate of records, which does not have diabetes and correctly classified as non-diabetic people. FP is the rate of records that are diabetics, however incorrectly classified as non-diabetics, and TN is the rate of non-diabetics that are classified as diabetics incorrectly. tenfold cross validation is used to train and tests the models.

4 Results

In this section, first, the dataset employed in this research is explained. Then, classification results achieved using the proposed method are presented.

4.1 Dataset

In this study, the PIMA Indian diabetes dataset is used to assess the proposed method for diagnosis Type 2 diabetes [26]. Insulin-glucose dynamics dictate the non-linearity of diabetes data [4], it is a non-linear dataset prepared from Indian women aged 21 years or older, and is available in the UCI's machine learning repository. It includes 768 records; each record is defined with eight integer-real attributes. There is another attribute, which is the label with values of 1 indicating patients that have diabetes and 0 for those that do not. Table 5 Definition of the Mathematical symbol used in equation (1) to (9)SymbolDescription racceptaccepting ratePpitchpitch adjustmentbrangepitch bandwidthrpaadjusting ratexoldcurrent pitch of a Harmony algorithmxnewpitch after adjustingEa random number between 0 and 1Prandomthe third component that is used to produce more variations of a solutionxidthe current position of a particlePidpbest of particlevidvelocity of particlePgdgbest of the groupWinteria factorc1relative influence of the cognitive componentc2relative influence of the social componentr1, r2random numbers used to keep the change of the population spread between 0 and 1, equallywmaxinitial weightwminfinal weightitermaxmaximum iteration numberItercurrent iteration number1 illustrate the attributes and their descriptions.

4.2 Results of the Proposed Algorithms

As presented at the beginning of Sect. 3, after feeding the PIMA Indian diabetes dataset into the system, three hybrids feature selection algorithms, including GA-Kmeans, GA-PSO-Kmeans, and HR-Kmeans are applied. The best features selected by HR-Kmeans, include BloodPressure, Glucose, and Insulin. Besides, Decision tree, SVM, and KNN are also examined on the same dataset. We divided the dataset into the train and the test set, 75% for train and 25% for test, respectively. The results of the different examinations mentioned are shown in Table 2.

We first examined the standard SVM, KNN, and DT, where 82.85%, 84.30%, and 86.63% accuracy were achieved without feature selection, respectively. Then, a different combinations of GA, PSO, Harmony, and K-means algorithm. Although the combination of GA and K-means produced better results than standard algorithms, more combinations

Table 1 Description of PIMA Indian diabetes records [11]

	Feature name	Feature description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test (mg/dl)
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin (mu U.ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Age (years)
9	Class label	Class variable (0 or 1)

Table 2 The results of the proposed hybrid methods and three standard classification methods on the PIMA Indian diabetes dataset

Feature selection	Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)
–	SVM	76.60	42.36	82.85
–	DT	81.31	75.33	84.30
–	KNN	88.27	93.13	86.63
GA	KNN	89.01	85.09	88.02
PSO	KNN	87.22	85.09	87.22
HR	KNN	90.15	88.02	90.55
GA-Kmeans	KNN	83.73	50.00	88.02
GA-PSO-Kmeans	KNN	86.65	75.33	89.64
HR-Kmeans	KNN	91.11	50.00	91.65

Table 3 The comparisons of the accuracy of different standard classifiers examined in this paper and reported in previous studies

References	Classifier	Accuracy (%)
[35]	Bagged tree	73.20
[35]	RUSBoosted trees	73.40
[35]	Boosted tree	75.00
[18]	C4.5	76.52
[18]	Naïve Bayes	76.96
[18]	LR	78.69
This paper	SVM	82.85
This paper	DT	84.30
This paper	KNN	86.63

are examined for better accuracy. Finally, the best accuracy of 91.65% was obtained using the HR-Kmeans hybrid.

4.3 Comparison the with Different Standard Algorithms

A large number of researches are reported on the classification of the PIMA Indian diabetes dataset. To show the superiority of the proposed algorithm, the results reported in some of the previous researches using the standard algorithm and without feature selection are provided in Table 3.

From the results reported in the above classifiers and standard classifiers examined in this paper, KNN achieved better results. Therefore, we tried to propose feature selections that could improve the results of KNN further. The results of the new combinations examined are provided in the next section.

4.4 Comparison with Different Hybrid Algorithms

Besides the standard algorithm, several combinations and hybrid of algorithms were reported in previous studies. Some of the reported researches, combined different classification algorithms to improve the overall accuracy. However, some of the researches added

the feature selection stage before the main classification stage and achieved better results based on the more discriminative features. A number of the researches with higher accuracies reported recently are presented in Table 4. All the presented results were based on the same dataset, PIMA Indian diabetes dataset. As shown in Table 4, some of the proposed methods have a feature selection stage in their classification process. However, they did not name the selected features. The name of the features selected by the method proposed in this paper are included in Table 4 for more accurate comparisons.

In three hybrid methods examined in this paper, Glucose, bloodPressure, and Insulin are selected in feature selection stage. In addition to the common features, in GA-Kmeans and GA-PSO-Kmeans combinations, Age and BMI are also selected. This is comparable by Age, BMI, and Glucose features selected in [26]. The selection of BloodPressure and Insulin features is noticeable from the results of the current research. These two features were not selected by the methods proposed.

Besides, the algorithms' processing time could be compared. However, the research studies reviewed in this paper did not report their processing time except the study presented by Kannadasan et al. [32]. In this paper, the best combination of algorithms was HR-Kmeans, where model processing time is 24.44 s. For GA-Kmeans and GA-PSO-Kmeans, model processing times are 22.3 s and 24.43 s, respectively. The model processing time for GA-PSO-Kmeans and HR-Kmeans are almost the same. However, HR-Kmeans achieved better accuracy. To compare the model processing time of the proposed method with that of the methods proposed before, a valid comparison could not be performed because experimental setups differ for further researches. Moreover, most of the previous studies did not report their model processing time. In the literature reviewed for this study, Kannadasan et al. [32] reported the model processing time, which in the best practice, it was 60.3 s.

As depicted in Table 4, the combination of metaheuristic algorithm, with the K-means clustering algorithm, produced an appropriate and successful feature selection method. All three combinations examined obtained better results than the best previously reported results. This approves the positive role of the K-means clustering algorithm in the feature selection stage. Because it could extract the hidden relationship between the records and assign them into one cluster. Then metaheuristic algorithms used such hidden relationships to rank the features for classification in the next stage.

Table 4 Comparison of hybrid algorithms proposed in previous researches with the hybrid algorithms proposed in this paper

References	Feature selection	Classifier	Features Selected	Accuracy (%)
[18]	PSO	Naïve Bayes	All 8 features	78.69
[18]	PCA	LR	All 8 features	79.56
[26]	BCO	Fuzzy	Age, BMI, Glucose	84.21
[33]	ANT FDCSM	Fuzzy rule miner	NA	87.7
[39]	SOMSwram	DNN	NA	80
[29]	Stacked-autoencoders SAE	DNN	NA	86.26
This paper	GA-Kmeans	KNN	Glucose, BloodPressure, Insulin, Age	88.02
This paper	GA-PSO-Kmeans	KNN	Glucose, BloodPressure, Insulin, BMI	89.64
This paper	HR-Kmeans	KNN	Glucose, BloodPressure, Insulin	91.65

Metaheuristic algorithms used in this paper, GA, PSO, and Harmony search algorithm, are population-based algorithms. GA and PSO are inspired by the swarming and collaborative behavior of the biological population [40]. Both algorithms depend on the exchange of information among members of the population, using deterministic and probabilistic rules to improve research results. However, PSO is computationally more efficient than GA [41].

The Harmony search algorithm is among the population-based metaheuristic algorithms as well. It has three parameters, including pitch adjustment rate, memory consideration rate, and memory size; however, in some applications the Harmony search algorithm is not sensitive to its parameters. Implementation of Harmony search algorithm is also easier than GA and PSO. In addition, multiple harmonies could be employed in parallel, resulting in higher efficiency [29]. Overall, the computational complexity and performances of the methods are application dependent. In the case of diagnosis diabetes based on the PIMA Indian diabetes dataset, the Harmony search algorithm achieved better accuracies than other metaheuristic algorithms and better model processing time than those of GA and PSO.

5 Conclusion

Individuals with diabetes are more likely to be infected by COVID-19. Therefore, providing a method that can predict or diagnose diabetes can decrease the mortality rate of COVID-19. In this paper, three feature selection algorithms are proposed for improving the accuracy of diabetes dataset classification. The PIMA Indian diabetes dataset is used. Different combination of metaheuristic algorithms, including GA, PSO, and the Harmony are examined in combination with the K-means clustering algorithm for feature selection. Then, KNN is employed for the classification of the diabetes records. The proposed combinations produced superior results than the results reported before on the same dataset, where the HR-Kmean hybrid gained the maximum accuracy of 91.65%. A future direction, the proposed algorithms could apply to local diabetes data. In addition, another combination of metaheuristic algorithms could be examined. Moreover, new fitness functions could be proposed for each of the heuristic algorithms for better feature rankings. The proposed method could be evaluated through mathematical and statistical tests such as McNamara's test in the future.

Appendix

Table 5.

Table 5 Definition of the Mathematical symbol used in Eq. (1) to (9)

Symbol	Description
r_{accept}	Accepting rate
P_{pitch}	Pitch adjustment
b_{range}	Pitch bandwidth
r_{pa}	Adjusting rate
x_{old}	Current pitch of a Harmony algorithm
x_{new}	Pitch after adjusting
E	A random number between 0 and 1
P_{random}	The third component that is used to produce more variations of a solution
x_{id}	The current position of a particle
P_{id}	p_{best} of particle
v_{id}	Velocity of particle
P_{gd}	g_{best} of the group
W	Interia factor
c_1	Relative influence of the cognitive component
c_2	Relative influence of the social component
r_1, r_2	Random numbers used to keep the change of the population spread between 0 and 1, equally
w_{max}	Initial weight
w_{min}	Final weight
$iter_{max}$	Maximum iteration number
Iter	Current iteration number

References

1. Mirza S, Mittal S, Zaman M (2018) Decision support predictive model for prognosis of diabetes using SMOTE and decision tree. Int J Appl Eng Res 13(11):9277–9282
2. Moreira LB, Namen AA (2018) A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. Comput Methods Programs Biomed 165:139–149
3. Ahmadi N (2020) Review of terrestrial and satellite networks based on machine learning techniques. J Soft Comput Decis Support Syst 7(3):13–22
4. Boiroux D, Aradóttir TB, Nørgaard K, Poulsen NK, Madsen H, Jørgensen JB (2017) An adaptive nonlinear basal-bolus calculator for patients with type 1 diabetes. J Diabetes Sci Technol 11(1):29–36
5. Favalli EG, Ingegnoli F, De Lucia O, Cincinelli G, Cimaz R, Caporali R (2020) COVID-19 infection and rheumatoid arthritis: faraway, so close! Autoimmun Rev 102:523
6. Muniyappa R, Gubbi S (2020) COVID-19 pandemic, coronaviruses, and diabetes mellitus. Am J Physiol Metab 318(5):E736–E741
7. Wiemken TL, Kelley RR (2019) Machine learning in epidemiology and health outcomes research. Annu Rev Public Health 41:21–36
8. Vaishya R, Javaid M, Khan IH, Haleem A (2020) Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr Clin Res Rev 20:20
9. Fang L, Karakiulakis G, Roth M (2020) Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet Respir Med 8(4):e21
10. Nilashi M, Samad S, Yadegaridehkordi E, Alizadeh A, Akbari E, Ibrahim O (2019) Early detection of diabetic retinopathy using ensemble learning approach. J Soft Comput Decis Support Syst 6(2):12–17
11. Sambyal N, Saini P, Syal R (2020) A review of statistical and machine learning techniques for microvascular complications in type 2 diabetes. Curr Diabetes Rev 2:19

12. Casalino G, Castellano G, Consiglio A, Liguori M, Nuzziello N, Primiceri D (2019) A predictive model for MicroRNA expressions in pediatric multiple sclerosis detection. In: International conference on modeling decisions for artificial intelligence, pp 177–188
13. Waring J, Lindvall C, Umeton R (2020) Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 104:101822
14. Guo W et al (2020) Diabetes is a risk factor for the progression and prognosis of COVID-19. Diabetes Metab Res Rev 2:e3319
15. Buettner R, Klenk F, Ebert M (2020) A systematic literature review of machine learning-based disease profiling and personalized treatment. In: 2020 IEEE 44th annual computers, software, and applications conference (COMPSAC), pp 1673–1678
16. Antosik-Wójcińska AZ et al (2020) Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling. Int J Med Inform 138:104131
17. Swapna G, Kp S, Vinayakumar R (2018) Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Comput Sci 132:1253–1262
18. Choubey DK, Kumar P, Tripathi S, Kumar S (2020) Performance evaluation of classification methods with PCA and PSO for diabetes. Netw Model Anal Heal Inf Bioinf 9(1):5
19. Dennis C, Engelbrecht AP, Ombuki-Berman BM (2020) An analysis of activation function saturation in particle swarm optimization trained neural networks. Neural Process Lett 52(2):1123–1153
20. P. V. S. Ganesh and P. Sripriya, "A Comparative Review of Prediction Methods for Pima Indians Diabetes Dataset," in *International Conference On Computational Vision and Bio Inspired Computing*, 2019, pp. 735–750.
21. Gujral S (2017) Early diabetes detection using machine learning: a review. Int J Innov Res Sci Technol 3(10):57–62
22. Fatima M, Pasha M (2017) Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl 9(01):1
23. Kandhasamy JP, Balamurali S (2015) Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci 47:45–51
24. Ertuğrul ÖF (2018) Two novel versions of randomized feed forward artificial neural networks: Stochastic and Pruned Stochastic. Neural Process Lett 48(1):481–516
25. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci 82:115–121
26. Beloufa F, Chikh MA (2013) Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Comput Methods Programs Biomed 112(1):92–103
27. Singh A, Gupta G (2019) ANT_FDCSM: A novel fuzzy rule miner derived from ant colony metaheuristic for diagnosis of diabetic patients. J Intell Fuzzy Syst 36(1):747–760
28. Laxmi S, Gupta SK (2020) Intuitionistic Fuzzy proximal support vector machines for pattern classification. Neural Process Lett 51(3):2701–2735
29. Hasan S, Shamsuddin SM (2019) Multi-strategy learning and deep harmony memory improvisation for self-organizing neurons. Soft Comput 23(1):285–303
30. Zhu J, Xie Q, Zheng K (2015) An improved early detection method of type-2 diabetes mellitus using multiple classifier system. Inf Sci (Ny) 292:1–14
31. Ramezani R, Maadi M, Khatami SM (2018) A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Eng J 57(3):1883–1891
32. Kannadasan K, Edla DR, Kuppili V (2019) Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clin Epidemiol Glob Heal 7(4):530–535
33. Choubey DK, Kumar M, Shukla V, Tripathi S, Dhandhania VK (2020) Comparative analysis of classification methods with PCA and LDA for diabetes. Curr Diabetes Rev 16(8):833–850
34. Al-Zebari A, Sengur A (2019) Performance comparison of machine learning techniques on diabetes disease detection. In: 2019 1st international informatics and software engineering conference (UBMYK), pp 1–4
35. Choudhary A, Kumar M, Gupta MK, Unune DK, Mia M (2019) Mathematical modeling and intelligent optimization of submerged arc welding process parameters using hybrid PSO-GA evolutionary algorithms. Neural Comput Appl 3:1–14
36. Apaza FV, Saire JEC (2018) Experimental test of harmony search algorithm with thousands of dimensions. In: 2018 IEEE sciences and humanities international research conference (SHIRCON), pp 1–5
37. Eberhart R, Kennedy J (1995) Particle swarm optimization. Proc IEEE Int Confer Neural Netw 4:1942–1948
38. Izzo D, Sprague CI, Tailor DV (2019) Machine learning and evolutionary techniques in interplanetary trajectory design. In: Modeling and optimization in space engineering, Springer, pp 191–210

39. Mohamadi M, Tab FA, Soltanian K (2019) Evolutionary feature selection based on semi-local search. In: 2019 4th international conference on pattern recognition and image analysis (IPRIA), pp 228–233
40. Souri A, Ghafour MY, Ahmed AM, Safara F, Yamini A, Hoseyninezhad M (2020) A new machine learning-based healthcare monitoring model for student's condition diagnosis in Internet of Things environment. Soft Comput 24:17111–17121
41. Souri A, Mohammed AS, Potrus MY, Malik MH, Safara F, Hosseinzadeh M (2020) Formal verification of a hybrid machine learning-based fault prediction model in Internet of Things applications. IEEE Access 8:23863–23874

Authors and Affiliations

Xiaohua Li[1] · Jusheng Zhang[2,3] · Fatemeh Safara[4]

Xiaohua Li
xhuali6@126.com

Fatemeh Safara
fsafara@yahoo.com

1 School of Physical Education, Hunan University of Arts and Science, Hunan 415000, China

2 South China University of Technology, Guangzhou 510641, Guangdong Province, China

3 Shenzhen Yiqizu Technology Co, Shenzhen 518000, Guangdong Province, China

4 Department of Computer Engineering, Islamshahr Branch, Islamic Azad University, Islamshahr, Iran