Soft Computing (2019) 23:9995–10017
https://doi.org/10.1007/s00500-018-3555-5

METHODOLOGIES AND APPLICATION

Assessment of the risk factors for type II diabetes using an improved combination of particle swarm optimization and decision trees by evaluation with Fisher's linear discriminant analysis

A. Sheik Abdullah[1] · S. Selvakumar[2]

Published online: 3 October 2018

Abstract

Type II diabetes is one of the chronic diseases, which is the cause of death and disability in most of the countries. The objective of this research work is to apply particle swarm optimization (PSO) algorithm with improved operating parameters along with decision tree (C4.5) for accessing the risk factors for type II diabetes. The model developed with the consideration of type II diabetic risk has parameter values with normal and abnormal levels. Experimental analysis has been carried out using an improved combination of PSO with decision trees and its various splitting measures with real-world diabetic dataset. The improvement in PSO is made by proposing self-adaptive inertial weight with modified convergence logic for the particles to accelerate in the given search space. From the result analysis, it has been observed that the risk factors corresponding to diabetes are postprandial plasma glucose (PPG), glycosylated hemoglobin (A1c), mean blood glucose (MBG) and fasting plasma glucose (FPG) identified with an improved accuracy more than that of the existing methods and algorithms. The efficiency of prediction has been tested using Fisher's linear discriminant analysis. From the inference, it has been observed that there occurs a strong relationship between the risk factors such as PPG, A1c, MBG and FPG with the risk corresponding to type II diabetes. Hence, predictive analytics using an improved combination of PSO with decision trees can also be deployed for the identification of risk factors corresponding to other chronic diseases such as coronary heart disease and kidney disease.

Keywords Data classification · Data mining · Data analytics · Decision trees · Discriminant analysis · Optimization · Prediction · Swarm intelligence · Type II diabetes

1 Introduction

Diabetes is a metabolic and chronic disease which is described by the prominent levels of blood glucose levels. Type II diabetes is the most common form of diabetes. With this form of diabetes, the body does not use insulin at proper levels. Hence, an insulin resistance is created which makes an oscillation over the insulin levels. At time intervals, the human body cannot make insulin at proper levels to keep the values of blood glucose at normal levels as depicted by Nathan et al. (2008). The prevalence of diabetes and its growth have been increasing enormously over all the regions in the world. The impact of diabetes caused about 1.5 million deaths in 2012, and higher values of blood glucose caused 2.2 million deaths with an increased risk of cardiac problems and other diseases. During this period, 80% of the disease occurred in low- and middle-age countries (Collen 1994). The World Health Organization (WHO) states that during 2030 diabetes will be one among the major causes of death over various regions in the world. In 2014, about 422 million adults (merely 8.5% of the total population) were affected due to diabetes as compared to 4.7% in 1980. The problem of diabetes may lead to stroke, heart attack, kidney failure, blindness and limp difficulty (Hapsara 2005).

According to a study conducted by the Indian Diabetes Foundation, in India, about 63 million people suffer from

Communicated by V. Loia.

✉ A. Sheik Abdullah
asait@tce.edu

S. Selvakumar
sselvakumar@ieee.org

[1] Department of Information Technology, Thiagarajar College of Engineering, Madurai 620015, Tamil Nadu, India

[2] Department of Computer Science and Engineering, GKM College of Engineering and Technology, Chennai, Tamil Nadu, India

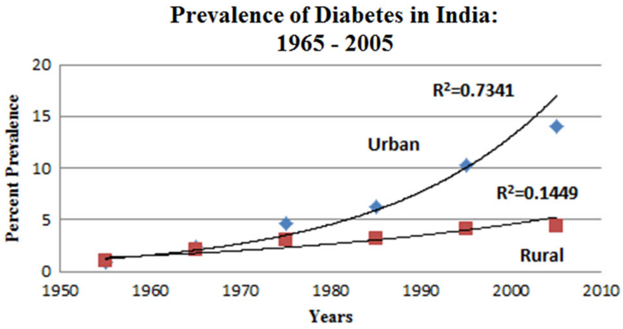


Fig. 1 Prevalence of diabetes in India 1965–2005

possible disability which is expected to rise up to 80 million during the year 2025. The survey made by Dr. Anoop Misra states that about 37% of the urban south Indians suffer from diabetes and prediabetes. The work by them provides a statistical report from the year of 1965–2005 which conveys segregation among the urban and rural population having diabetes. The report states that there are a significant increase in urban area with an exponential trend $R^2 = 0.74$ and a slower range in the rural area of $R^2 = 0.28$. The statistics reveal that the ratio of the population with diabetes varies from 5.4% in the northern region, 12.3–15.5% in Chennai zone to a range of 16.8% in central India as stated by Gupta and Misra (2007). The prevalence of diabetes is depicted in Fig. 1.

The intersection among medical science and information technology provides a solution toward data-driven decision making. The future community is in need of an integrated access to clinical information to formulate computer-based data management and decision-making process. Hence, it provides the way not only for managing electronic health records which engage clinical and patient information but also for obtaining a data model which helps in deciding proper levels as illustrated by Shortliffe and Cimino (2014). The process of decision making includes major components such as the data elements (attributes), measurement levels, stopping criteria, test process and the implications of the attribute level with its corresponding process. Population-based healthcare study will provide the significant features that are more predominant toward the disease. The significance of disease-specific syndromes with the location of the people could solve the problems of the origins and progress of the disease over the regions (Anon 2016). The notable benefits with healthcare data analytics and feature selection include the following:

1. Reducing clinical cost.
2. Providing a decision support model.
3. Risk determination with specific reference to locality, likelihood and dietary habits.
4. Healthcare coordination.
5. Improvement in patient monitoring systems.

Different sorts of prediction models have been deployed in the medical field which serve as the backbone for disease management and diagnosis (Steyerberg 2009). Most of the techniques related to supervised learning techniques were in practice for clinical prediction. Some of them include linear regression model, logistic regression, neural network, decision trees and survival data models. These techniques focus on the significance and nature of the data which then provide the relationship among the features and between the features over the target variable. Among the various models, choosing the suitable one for the particular healthcare problem depends upon the nature of the data and the relationship that exists between the attributes. There are various forms of algorithmic methods that exist, but some focus only on feature selection and only some of the methods on data classification. Algorithms used in the existing system are multi-agent PSO, genetic algorithm, support vector machine classifiers, k-means clustering, logistic regression, decision trees, Naïve Bayes, sensitivity and specificity measures. These methods have been used explicitly for either classification or prediction. The major limitations observed are as follows:

1. There seems to be no form of improvement in the operating parameters of the existing algorithms.
2. Most of the methods have been dealt with benchmark dataset rather than focusing on real-world data.
3. The application of feature selection and classification methods for real-world dataset in medical informatics lags in accordance with risk determination and analysis.

The proposed research work focuses upon the development of an algorithmic method using PSO with its improvement in operating parameters with decision tree algorithm. Data corresponding to type II diabetes from a regional hospital were used in developing the algorithmic model. The operating characteristics of PSO algorithm depend upon the fixing of inertial weight in accordance with the number of iterations and population generation. The efficiency of the PSO algorithm depends upon the phenomenon of fixing up of inertial weight in accordance with particles position and its velocity. In order to make the particles accelerate for an optimal solution and to improve the comprehensibility of the model in the given search space, we have proposed a scheme for fixing a self-adaptive inertial weight with modified convergence logic. This mechanism makes the particle accelerate in its logical space to show a variation in each set of iteration at individual levels. With this, the movement of the particle is assigned with a small or larger value of inertial weight.

Once the features are selected, the fitness function is evaluated with the total number of positive and negative cases which have been predicted by the decision tree algorithm. The efficiency of the method is tested with Fisher's linear discriminant analysis with test case interpretation and evaluation. The correlation between the selected features is determined using curve fitting by an exponential distribution using R^2 analysis. Section 2 provides the literature about the existing models; Sect. 3 describes the proposed work; Sect. 4 includes the experimental results and discussion, and Sect. 5 concludes the observed results with a set of selected features, accuracy and R^2 values for dependency among the selected features. The future work can be related to the risk factor determination of various other non-communicable diseases.

2 Literature review

The mechanism of data classification algorithms in lung cancer prediction plays a substantial role in risk determination. Support vector machine classifier is mainly suitable for pattern classification problems. The work by the authors Xue et al. (2014) by using support vector machine classifiers with feature selection process provides an improved accuracy with the expected predictive measure. The observed optimal solution with the combination of swarm intelligence techniques with classification suits well for pattern identification and dissemination in lung cancer dataset. Hence, a theoretical approach has also been given to parameter fixing and convergence for PSO. Similarly, the work by the authors Chen et al. (2014) over gene expression microarray data using correlation-based feature selection and a variant of PSO delivered showed the decrease in the error rate. Gene selection and distinguished class variants also seem to be significant over the iteration stages at all levels. Similar work has been carried out for cancer microarray data for feature selection as illustrated by Sahu and Mishra (2012). The result provided an improved accuracy with better prediction, but a limitation was observed with the convergence speed of particles upon generation at each level.

An experimental and theoretical approach has been given by the authors Tang et al. (2015). It states that PSO well suits all domains of problems irrespective of the nature of the data and their constraints. Meanwhile, the authors have explained all strategies related to the objective function, convergence characteristics, selection parameters and discrete-valued functions which are the most possible applications of the PSO algorithm. The experimental results have also proved that PSO outperforms other swarm intelligence techniques over various datasets.

The work proposed by Lee and Kim (2016) investigated the phenotype hypertriglyceridemia waist which has a strong correlation with type II diabetes. They have assessed the predictive power concerned with phenotypes and triglyceride levels. The data correspond to Korean adults for the cross-sectional study. The interpretation has been made using naïve Bayes algorithm and logistic regression for predicting the power of various phenotypes. The evaluation has been performed using a tenfold cross-validation confirming the presence of HW which has been strongly correlated. The predictive power confirms that women are more prone to type II diabetes than men during analysis. The findings made by the authors provided a clinical decision support system for screening at the initial stages of heart disease which seems to be a limitation of the model developed.

Data records corresponding to coronary heart disease (CHD) were collected at Paphos General Hospital in Europe by Karaolis et al. (2010). The data record corresponds to the following events such as myocardial infraction, coronary intervention and bypass graft surgery. The record has been segregated into before and after the event with the differentiation among the risk factors. The implementation was preceded with decision trees with an accuracy of about 66%, 75% and 75% for the events such as myocardial infraction, coronary intervention and bypass graft surgery, respectively. The authors used only decision trees for evaluating the risk corresponding to CHD regarding accuracy. Though even the authors used real-world medical data corresponding to CHD, they have not discussed the nature of risk factors, its correlation and dependencies.

The work with the development of a predictive model by Low et al. (2017) for kidney disease in progression to type II diabetes with data from a diabetic center in a regional hospital was monitored across eGFR upon various categories. The implementation proceeded with the random split among the tuples with 70% and 30% folds, respectively, for validation. The model used was a stepwise multivariate logistic regression with 42.9% progression toward kidney disease. From the results, it has been observed that for development and validation 42.2% and 44.6% have progressions toward type II diabetes. The ROC tests also incurred 0.80 and 0.83, respectively. Hence, with the clinical measurements, predictive models can be stimulated upon different algorithms for the determination of risk factors among diseases. The authors have claimed the use of logistic regression used to predict the disease, but the performance of logistic regression seems to be good when the dataset is as large as possible. Well-observed predictors and likely worthy ratio of positive and negative cases are illustrated by the authors Reed and Wu (2013).

Meanwhile, the examination of longitudinal data for the patients with type II diabetes with various risk factors has been analyzed using multivariate models by Ievers-Landis et al. (2015). The observation over a 2-year stipulated period provides 10–18% corresponding to smoking and drinking. The ethnicity and depression were also significant factors for the people over the regions. The predominant symptoms, stressful events and BMI were found to be the significant factors which have been considered over time.

The effect of non-communicable diseases has turned out to be the most adverse phenomenon over the regions of sub-Saharan Africa by West et al. (2016). The authors have outlined a screening algorithm for determination of the health conditions over the people. Medical cases of about 713 were considered from the general population for analysis. Sensitivity and specificity were used for optimizing Monte Carlo and hypercube model. Experimental results show that fasting blood glucose has been found to be the most predominant one with sensitivity and specificity of about 68% and 99%, respectively. The cost per patient was estimated to be \$2.94 and \$6.04 for HBA1c in accordance with the severe cases for type II diabetes. The authors sketched the model regarding sensitivity and specificity not through any data classification techniques. Also, the risk factors for the disease explored only the cost per patient corresponding to type II diabetes.

A research study with 586 consecutive patients with type II diabetes and coronary heart disease was observed. Experimental techniques such as sensitivity and specificity were applied to the data records. The results showed that AUC was 0.88 and sensitivity and specificity were 84% and 82% with positive for major and minor depression cases, respectively. Hence, this can be used to observe cases with major and minor depressions at optimal rates as illustrated by Van der Zwaan et al. (2016).

Diabetes gives rise to major complications such as blindness, heart disease, kidney disease and stroke in the region of USA by Khan et al. (2017). The authors applied bootstrap sampling method to detect the predictive bound over the diabetic data records. The authors used a random sample of data collected from National Health and Nutritional Examination Survey. Logistic regression was used to determine the significant risk factors with a p value <0.001 for age, total protein, food (fasting) and HDL. Hence, the maximum probable accuracy was obtained with the determination of extreme and mild values of outliers in type II diabetic prediction. Machine learning techniques were widely used to evaluate the risk related to medical data. The decision framework for prediction was proposed with machine learning models such as naïve Bayes, random forest, decision trees, SVM and logistic regression analysis by authors Zheng et al. (2017). Data records of 300 patients were used for analysis, and the results showed that AUC was achieved at the rate of ~0.98 on an average when compared with the existing methods.

Hybrid prediction model plays a significant role in data analysis. The work by the authors proposed a hybrid model for the prediction of type II diabetes using classical k-means clustering and C4.5 decision tree algorithm. Data corresponding to Pima Indian Diabetes obtained from UCI repository are used for analysis with k-fold cross-validation. The proposed model gives an accuracy of about 92.38% with an improved sensitivity and specificity measures over the observed data. In compliance with future work, real-world medical data can be used in prediction of risk concerning type II diabetes by Patil et al. (2010).

Similar work has been made for the determination of missing values over medical data by Purwar and Singh (2015). They developed a hybrid model for medical data prediction and evaluation. The observed results show that the incorporation of a hybrid model can be more preferably used for medical data analysis. The same strategy was adopted for heart disease prediction over medical data by Yang and Garibaldi (2015).

The outcome imposed a hybrid model for heart disease prediction. They used the rule-based technique with natural language processing (NLP) over medical data. The result shows that development of hybrid model provides an F-measure of 0.927 to the best over model evaluation. Table 1 describes the accuracy level, data used and model evaluated by various authors.

Based on the limitations and drawbacks of the existing techniques about algorithmic model and utilization of real-world dataset, there is a need for an optimal algorithm which encompasses feature selection and data classification with an improvement toward convergence speed, accuracy and execution time. In order to overcome the drawbacks, we have proposed an improved combination of PSO algorithm with decision trees for the assessment of risk factors that correspond to type II diabetes.

3 Proposed methodology using particle swarm optimization with J48 (Java implementation of C4.5 decision tree algorithm)

3.1 Rationale behind the proposed approach

Data optimization methods are deployed in various fields such as engineering, medical sciences, management, physical sciences and social learning. The process behind optimization is to choose the best solution from the set of feasible solutions, thereby providing scientific decision making. The problem formulation in optimization includes identifying the decision variables, objective function and learning factors which correspond to the analysis of the feature selection process. The next is to choose an appropriate numerical method to solve, test and make proper decision for the optimal solution obtained. There are various optimization methods, and some of them are enumeration methods, gradient methods, random search methods and meta-heuristic methods. Table 2 presents the summary corresponding to the limitations and usage of the optimization methods.

The advantage of meta-heuristic method behind enumeration, gradient-based and random search methods is: it

Table 1 Literature survey over medical data by various authors

S. no	References	Disease specified	Algorithm applied	Dataset used	Accuracy level
1.	Karaolis et al. 2010	Coronary heart disease	Decision trees	Real-world diabetic dataset (European People)	75.00%
2.	Patil et al. 2010	Diabetes	K-means clustering with C4.5 decision trees	Pima Indians diabetes dataset	92.38%
3.	Sahu and Mishra 2012	Cancer	PSO with SVM, KNN	Benchmark cancer dataset	96.00%
4.	Xue et al. 2014	Lung	PSO with SVM	Lung UCI	78.40%
5.	Tang et al. 2015	Non-communicable	MAPSO	Benchmark data	92.25%
6.	Lee and Kim 2016	Type II diabetes	Naïve Bayes with logistic regression	Real-world diabetic dataset (Korean people)	73.50%
7.	West et al. 2016	Type II diabetes	Sensitivity and specificity with Monte Carlo and hypercube model	Real-world diabetic data from rural Tanzania	–
8.	Low et al. 2017	Kidney disease with Type II diabetes	Stepwise multi-variable logistic regression	Real-world diabetic data from a regional hospital	75.60%

Table 2 Comparison among optimization methods

S. no.	Method	Limitations	Usage
1.	Enumeration methods	It requires high computation cost and provides finite feasible solution	Generally used to solve integer programming and combinatorial problems
2.	Gradient methods	These methods belongs to local search techniques and provides only local optimal solution	It is used to fix the search space which is proportional to the positive or negative gradient of the function
3.	Random search methods	These methods can be used in cases when the objective function is not continuous or differentiable	Easy to apply for complex problems without gradient information
4.	Meta-heuristic methods	Meta-heuristic algorithms perform differently in different types of applications. Statistical data analysis will justify its performance	Meta-heuristic methods can be used as common framework which can be applied to different problems with making modifications to adapt to the specific problem

corresponds to an iterative process with subordinate heuristic with a combination of different intellectual concepts for exploring and exploiting the given search space as given by Talbi (2009). These methods are stochastic in nature with random search corresponding to local and global best solution for the given search space as illustrated by Gendreau and Potvin (2010). There are various meta-heuristic methods such as genetic algorithm, evolution strategy, tabu search, differential evolution and swarm intelligence methods. When compared with other methods, the major advantage of swarm intelligence technique is: these methods have the ability to act in a coordinated way without the presence of a controller (coordinator). The applications of swarm intelligence focus on the behavior of ants, flocking of birds, swarm-based network and the behavior of wasps.

Among all the techniques in the swarm intelligence, PSO is a population-based intelligent system that requires initial population corresponding to random solutions. The search for the optimal solution is obtained by updating the particle generations without any external evolution operators such as selection, crossover and mutation. The particles in PSO fly over the given search space and learn accordingly with their own experiences with the velocity which is then provided with cognition factor, social factor and inertial weight. The fixing up of target value for inertial weight plays a major role in the exploration of the search space to find a near-optimal solution, which is the major contribution in this research work.

From the literature, it is noted that researchers in medical informatics focused either feature selection or data classification in medical data prediction. Also, the utilization of statistical measurements is limited to sensitivity, specificity, F-measure and ROC analysis. Hence, there should be an algorithmic method which focuses on feature selection, classification and statistical analysis in medical data for exploring the risk factors that contribute to the disease. Data classification and prediction play a significant role in risk factor determination and its relationship. Among all the classification techniques, decision tree has the unique ability to assign definite values to a given problem, decision with the outcome of reducing ambiguity in the decision-making process. The following are the specific characteristics that correspond to the selection of decision tree for our algorithmic model development with an improved combination of PSO algorithm.

- Specificity
- Comprehensive nature
- Transparency
- Flexibility
- Resilience and
- Data validation

With the specific characteristics, the decision tree provides a framework in order to quantify the observed values and probabilistic function of each possible outcome of a decision. This phenomenon makes data solvers and decision makers to have a good choice in data classification with that of available alternatives.

3.2 Description of the algorithm

The proposed model encompasses the procedure of feature selection paradigm using PSO with J48 (Java implementation of C4.5 decision tree algorithm). PSO is one of the biologically inspired algorithms that are stimulated by the behavior of bird flocking as illustrated by Kennedy and Eberhart (1995). PSO is a computational procedure which fundamentally centers toward optimizing a given problem in an iterative way to infer a sensibly ideal arrangement. The generated particles are accelerated with an initial velocity. The movement of each of the particles is determined by the particles *pbest*. The assurance of the best particle over the signified fitness function is dictated by the *gbest* value. The entire search space is updated by the position obtained by each particle over the iterations (Abdel-Kader 2010). The particles are accelerated to reach the optimal solution until the stopping criterion is reached as examined by Zyout et al. (2015). In general, PSO is a meta-heuristic algorithm which targets the determination of optimal solution for a given measure of quality. It does not use the derivatives to the selection of functions to the defined variables. The basic concept of PSO lies in accelerating each particle toward its *pbest* and the *gbest* locations, with a random weighted acceleration at each time step. The search space exploration is illustrated in Fig. 2. Each particle tries to modify its position using the following information (Abdel-Kader 2010):

1. The current positions.
2. The current velocities.
3. The distance between the current position and *pbest.*
4. The distance between the current position and the *gbest.*

The modification of the particle's position during iteration can be mathematically modeled according to Eq. 1:

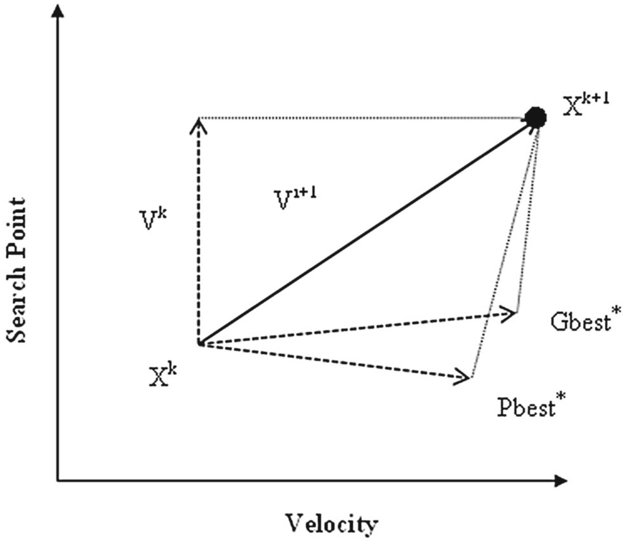


Fig. 2 PSO search space in an environment

$$V_i^{k+1} = wV_i^k + c_1 Rand_1 \times (pbest_i - s_i^k) + c_2 Rand_2 \times (gbest - s_i^k), \tag{1}$$

where V_i^{k+1}: weighting factor with new velocity at an iteration $k+1$, V_i^k: weighting factor with current velocity at an iteration k, w: inertial weighting function, $Rand$: the random number which is uniformly distributed between the ranges of [0–1], s_i^k: current position of the particle for the iteration k, $pbest_i$: the best position reached by the particle I, $gbest$: the global best position observed from the group of particles for the given iteration and c_1, c_2: learning factors.

The weighting function for the exploration of the particles needs to be determined in accordance with the particle's initial and final position. The mechanism of inertial weight follows a global search strategy and smaller inertial weight with local search strategy as illustrated by Shi and Eberhart (1998a). The size of the group for the set of particles has to be assigned which is defined by the factor N. Then, the generation of the initial population has to be made which ranges from $x_1, x_2, x_3 \ldots x_N$, the particles are accelerated with an initial velocity to reach its first new position, and this new position is given as: $x_1(1), x_2(1), x_3(1) \ldots x_N(1)$. The corresponding vector values are $x_i(1) = \{1, 2, 3 \ldots N\}$ called to be the vector coordinates of the particle. The objective function for the particles is given by: $f[x_1(1)], f[x_2(1)], f[x_3(1)] \ldots f[x_N(1)]$; the observed particles move accordingly to obtain a merely optimal solution with the triggered velocity updated at each level of iteration Chuang et al. (2011). The operation of the particle position can be made by Eq. 2 as:

$$P_i^{k+1} = P_i^k + V_i^{k+1}, \tag{2}$$

where P_i^k: current position of the particle i at iteration k and V_i^{k+1}: the updated velocity of the particle i at iteration k.

During iteration at each level, the particle finds its best position over the entire set of movement, and hence, it is referred to as the particles *pbest* value. Among the overall particles movements and its iterations, the observed best value (the best particle's position) is referred to as the *gbest* value. The values c_1 and c_2 are the cognitive learning factors which influence the social group. If the observed solution is convergent, then the iteration gets stopped Liu et al. (2011). The determination of the convergence at a point of contact can be ascertained by the following conditions:

1. The termination is made when maximum number of iterations is reached as specified.
2. Once the solution is acceptable, the iterations can be terminated.
3. If there is no improvement in the generation of swarms, then the process can be terminated.
4. If the swarm radius is closer to zero, the condition gets terminated.
5. If the slope of the objective function is close to zero, then the condition gets terminated.

The new position which has been determined must be away from the *gbest* and *pbest* positions. Only then the velocity of the particle ranges widely to a high value. The set of selected features is fed as an input to the J48. The J48 algorithm classifies the input features based upon the splitting criterion which has been used for evaluation in the determination of the accuracy over the selected set of features (Han 1996). The algorithm proceeds following the input parameters:

1. Training data partition (D).
2. The attribute list.
3. Attribute splitting criterion which determines the best splitting partition.

There are about various forms of attribute selection measures such as information gain and Gini index. If the data records are segregated in a binary format, then we can use Gini index as an attribute selection measure. But for our dataset, the values for each of the attributes seem to be continuous and discrete rather than binary value. Therefore, the information gain is used for the best splitting criterion in creating the tree. The information to classify a tuple is represented in Eq. 3 as:

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i). \tag{3}$$

The information to be obtained after the partition is given by Eq. 4 as:

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}\big(D_j\big). \tag{4}$$

The recursive partitioning ends when any one of the following conditions is reached:

1. If all the tuples of record belong to the same class in D.
2. If there occur no attributes for further partition.
3. If there occur no tuples for an identified branch during partition.

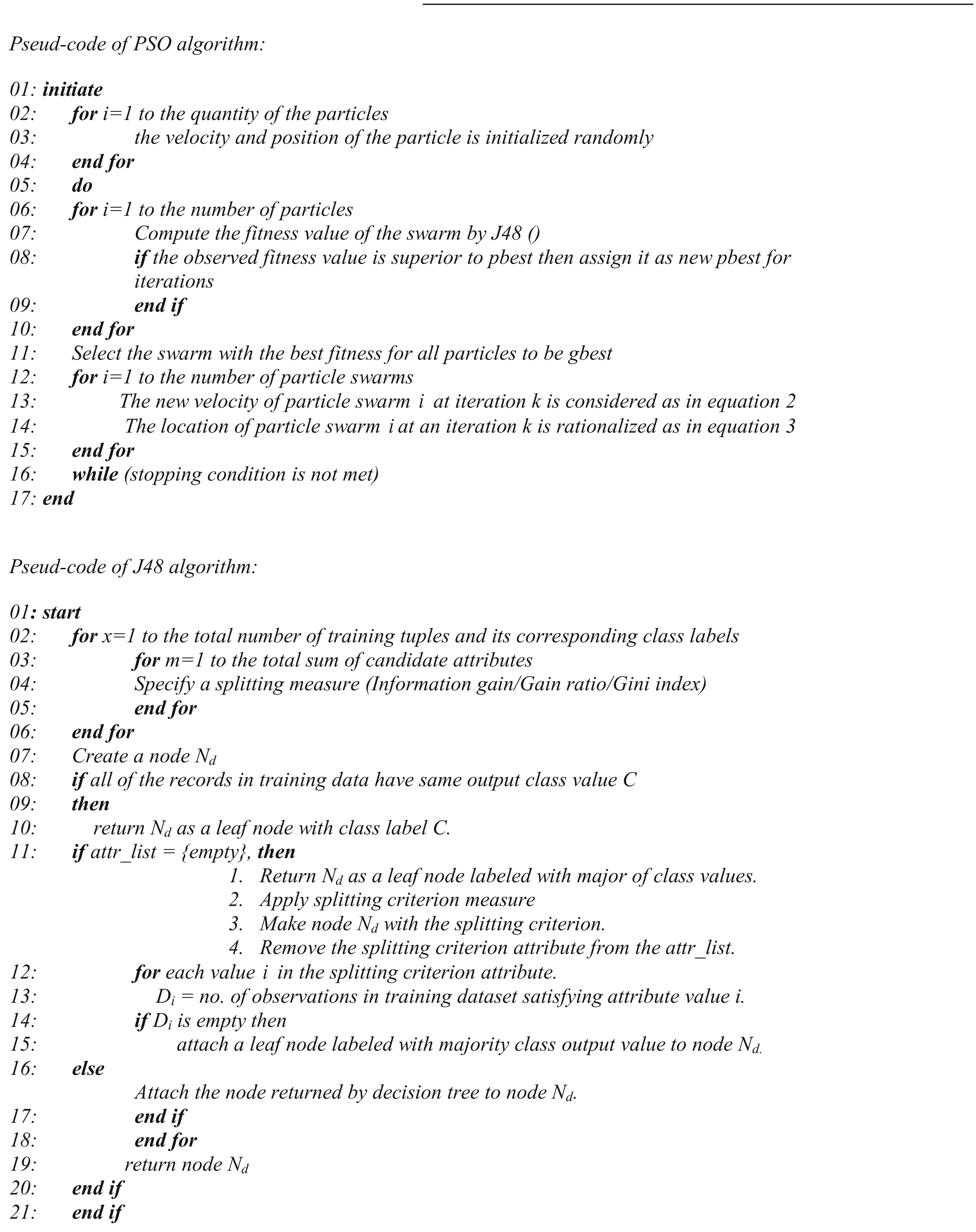
With the generated set of particles for each of the iteration, the subset of features is validated for its functional value in which each set is used up for training and testing over J48 classifier. Therefore, the fitness for each of the particles is evaluated. During iteration for the particles, the best fitness is the local best fitness value which is used to be the current fitness for the particle generated over time. The local best position for the first iteration is the current position for the particle if the local best position of that particle is nearer to the convergence point in the search space when compared with the previous stage then the position gets updated to the highest value over further iterations.

The global fitness value is then the maximum of the local best value observed so far for each set of iteration is concerned. The process continues until the convergence in the search space is reached. During iteration, the position and the velocity of the particles get updated simultaneously. The fitness is evaluated using J48 classifier under the following conditions:

1. If the current fitness is lesser than that of the local best fitness value.
2. If the current fitness is lesser than that of the global best fitness value.

Then, the global best fitness is set to be the current fitness or else the position and velocity of the particle get updated each time. At last when the stopping criterion is met, the procedure ends. With the number of tuples classified, if the class label of the nth tuple of the records is the same as that of the class label of the positive test case of the tuple, then the value of true positive gets incremented accordingly. Similarly, if the test cases of the negative tuple match with those of the classified tuple, then the value of the negative tuple gets incremented. The fitness value with the addition of correctly classified positive and negative tuples to that of the total number of records. The following pseudo-code provides the PSO algorithmic process in locating the best solution for the given search space.

3.3 Pseudo-code of PSO–J48



Pseud-code of PSO algorithm:

```
initiate
    for i=1 to the quantity of the particles
        the velocity and position of the particle is initialized randomly
    end for
    do
    for i=1 to the number of particles
        Compute the fitness value of the swarm by J48 ()
        if the observed fitness value is superior to pbest then assign it as new pbest for
        iterations
        end if
    end for
    Select the swarm with the best fitness for all particles to be gbest
    for i=1 to the number of particle swarms
        The new velocity of particle swarm i at iteration k is considered as in equation 2
        The location of particle swarm i at an iteration k is rationalized as in equation 3
    end for
    while (stopping condition is not met)
end
```

Pseud-code of J48 algorithm:

```
start
    for x=1 to the total number of training tuples and its corresponding class labels
        for m=1 to the total sum of candidate attributes
        Specify a splitting measure (Information gain/Gain ratio/Gini index)
        end for
    end for
    Create a node $N_d$
    if all of the records in training data have same output class value C
    then
      return $N_d$ as a leaf node with class label C.
    if attr_list = {empty}, then
                1. Return $N_d$ as a leaf node labeled with major of class values.
                2. Apply splitting criterion measure
                3. Make node $N_d$ with the splitting criterion.
                4. Remove the splitting criterion attribute from the attr_list.
        for each value i in the splitting criterion attribute.
          $D_i$ = no. of observations in training dataset satisfying attribute value i.
        if $D_i$ is empty then
            attach a leaf node labeled with majority class output value to node $N_d$.
    else
        Attach the node returned by decision tree to node $N_d$.
        end if
        end for
       return node $N_d$
    end if
    end if
```

3.4 Proposed workflow

The modules corresponding to the proposed workflow are data collection, data preprocessing, feature selection, classification and model evaluation. Data collection involves the retrieval of dataset related to type II diabetes such as patient's data, blood sample data and its comorbidities. Then, by utilizing the data analysis, the valuable information (patterns) from the dataset can be retrieved. The collected diabetes dataset corresponding to various sorts of events such as prediabetes, type I diabetes and type II diabetes will be segregated by assigning class labels to each of the corresponding patient records. From the set of records, the optimized feature set will be identified by applying PSO with proposed self-adaptive inertial weight by modified convergence logic. With the selected set of features, the fitness function is calibrated using J48 algorithm.

Among the variants of the decision tree classifier, the J48 has been found to be an improved one from the experiments made by Sheik Abdullah et al. (2017). For the initial phase of the iteration process, the particle's local fitness has to be estimated which, then, is assigned as the current fitness of the particle. Among all the observed local positions, the best position is chosen to be the current position of the particle as suggested by Lin et al. (2008). At each level, in updating the particle position and velocity values, the fitness of the particle is evaluated using the J48 classifier. The workflow of the proposed model is illustrated in Fig. 3. The set of classified tuples of records will then be observed from the confusion matrix with true positive, true negative, false positive and false negative. The grouping among the set of positive and negative tuples of records is calculated as per the process described in Fig. 4.

With the generated set of particles for each of the iteration, the subset of features is validated for its functional value, in which each set is used up for training and testing over J48 classifier as illustrated by (Sheik Abdullah 2012). Therefore, the fitness for each of the set will be evaluated. During iteration over the set of particles, the best fitness is the local best fitness value which is used to be the current fitness for the particle generated over time. The local best position for the first iteration is the current position for the particle if the local best position of that particle is nearer to the convergence point in the search space. When compared with the previous stage, the position gets updated to the highest one over further iterations. The global fitness value is, then, the maximum of the local best value observed so far for each set of iterations. The process gets continued until the convergence in the search space is found to be elevated. For each of the iterations, the position and the velocity of the particles get updated simultaneously. The fitness is evaluated using J48 classifier under the condition if the current fitness is lesser than that of local best fitness and also the current fitness is lesser than the global best fitness. Then, the global best fitness is set to be the current fitness or else the position and the velocity of the particle get updated each time. At last when the stopping criterion is met, the procedure ends.

4 Experimental results and discussion

4.1 Learning population

A total of 732 subjects were collected for the period of 2012–2015 under the examination of a medical expert. Written informed permission was obtained from the medical experts for the medical study. The subjected data have been preprocessed according to the mechanism of data cleaning, identifying duplicates, and the data are made suitable for processing.

To handle the numeric data values appropriate to data analysis, the values have been signified to the min–max range with zero mean or Z-score data standardization technique. In this method, the working principle is given in Eq. 5 as:

$$Z = \frac{X_i - \mu}{\sigma}, \tag{5}$$

where X_i—value of the attribute A, μ—mean value and σ—standard deviation.

The observed data contain the data attributes corresponding to the comorbidities, informative patterns, and prediabetic nature corresponding to type II diabetes. From the set of observed records, optimized feature set will be determined using PSO and C4.5 decision tree algorithm (J48). Finally, the data tuples and their associated class labels with optimized features get examined and analyzed by medical practitioners. With the observed model, the existence of type II diabetes and its risk can be evaluated by medical experts.

4.2 Attribute description and measurement

The processed medical data have 23 attributes including the class label as described in Table 3. Each of the defined attributes has its exact specification level of normal, prediabetics condition and the occurrence of diabetes.

The range values for each of the attributes vary significantly for male and female subjects who are prone to type II diabetes as suggested by Nathan et al. (2008). The ascertained values have been signified according to the Diabetic Association of India and World Health Organization (WHO).

4.3 PSO operating parameters and learning factors

In the development of the prediction model using PSO–J48, the execution of PSO for the defined number of iterations

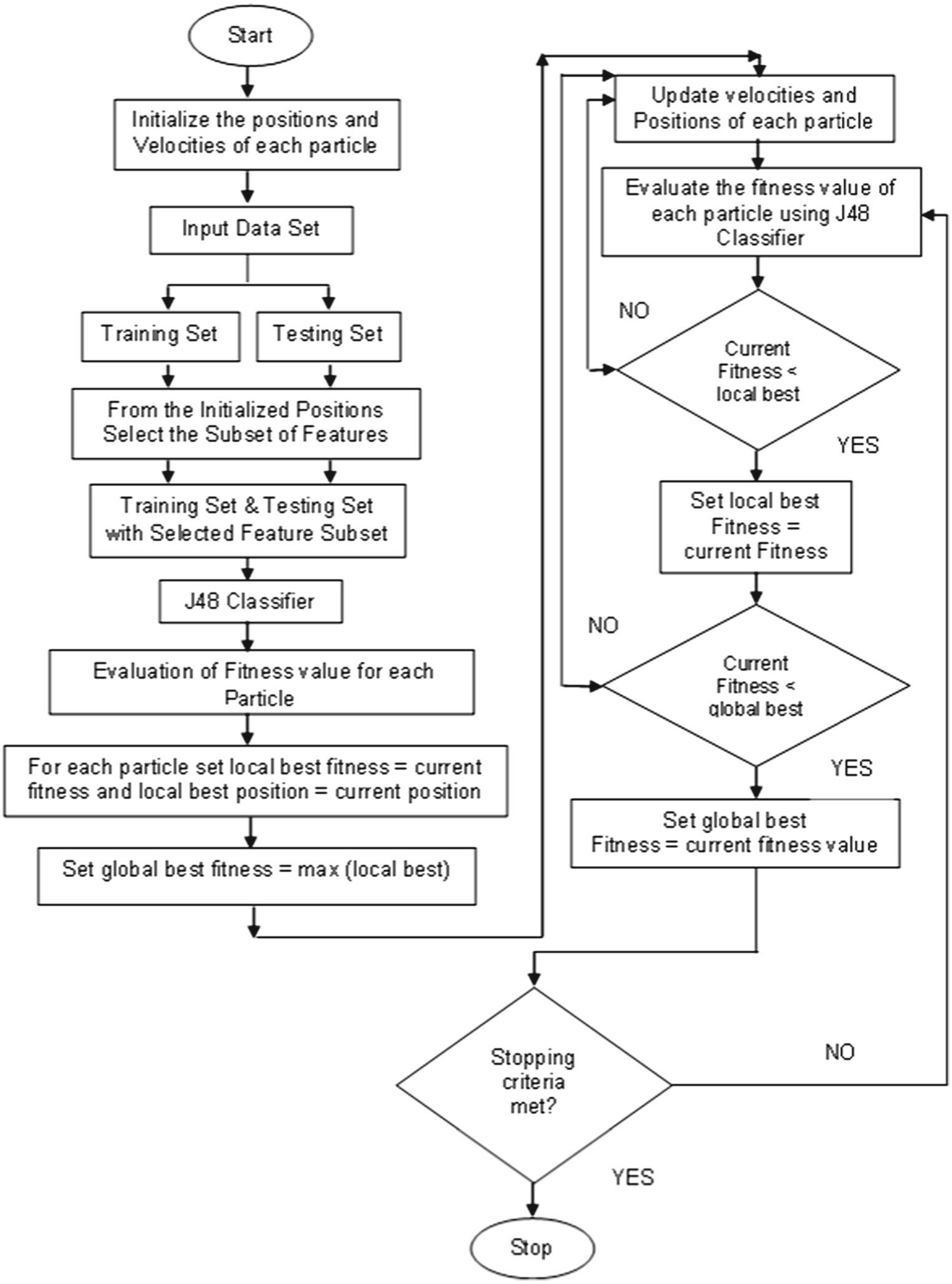

Start
Initialize the positions and Velocities of each particle
Input Data Set
Training Set
Testing Set
From the Initialized Positions Select the Subset of Features
Training Set & Testing Set with Selected Feature Subset
J48 Classifier
Evaluation of Fitness value for each Particle
For each particle set local best fitness = current fitness and local best position = current position
Set global best fitness = max (local best)
Update velocities and Positions of each particle
Evaluate the fitness value of each particle using J48 Classifier
NO
Current Fitness < local best
YES
Set local best Fitness = current Fitness
NO
Current Fitness < global best
YES
Set global best Fitness = current fitness value
Stopping criteria met?
NO
YES
Stop

Fig. 3 Proposed methodological workflow using PSO–J48

needs, the parameter has to be definite. There are about two learning factors C_1 and C_2 (cognition and social learning factors) which have to be fixed during the calculation of the fitness function. The cognition factor is defined as $C_1 r_{i,n}^j \left(P_{i,n}^j - X_{i,n}^j \right)$ in which it represents the own particles experience. The social learning factor then reflects the evidence shared by all the particles as expressed by $C_2 r_{i,n}^j \left(G_{i,n}^j - X_{i,n}^j \right)$. The learning factors play a significant role in exploring the particle in the given search space as illustrated by Eberhart and Shi (2001).

If the values of C_1 and $C_2 = 0$, then there exists no cognition and social learning for the particles about velocity updates. Therefore, we cannot observe an optimal solution unless the solution is on its path of observance. Upon the condition for C_1 to be 0, the consequence for velocity update is known to be the 'social' only learning model. By having an interaction between the particles, the PSO has its ability to formulate into a new search space. Due to this behavior, the

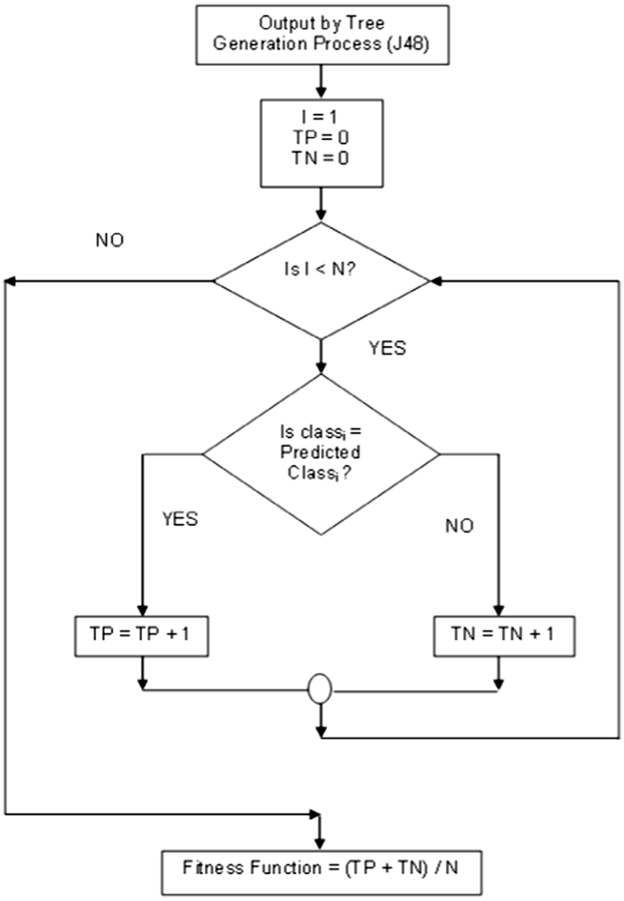

Output by Tree Generation Process (J48)
I = 1
TP = 0
TN = 0
NO
Is I < N?
YES
Is class$_i$ = Predicted Class$_i$?
YES
NO
TP = TP + 1
TN = TN + 1
Fitness Function = (TP + TN) / N

Fig. 4 Computation of the fitness value over the classified records

particles converge to a faster speed of action. The solution might be successful for some of the test cases upon performance analysis.

If $C_2 = 0$, then there occurs no social learning among the particles about velocity update. Hence, the search for the best position will not be shared as information among the particles. Therefore, unless the information has been shared among the particles across the individuals, each one of the particles is said to run independently. Upon these conditions, the algorithm is said to have slower convergence rate than that of the previous one as similar to the experiments by García-Nieto and Alba (2010). About the learning factors C_1 and C_2, the PSO has its operating parameter which is said to be the inertial weight as illustrated by Wang et al. (2017). If the inertial weight for the particles is not fed for the particles in searching the best position, then the velocity of the particles is said to be memory-less with the convergence factor more rapidly than that of the original form of PSO. Hence, without inertial weight, the particles have their ability to sample their positions in and around the *pbest* and *gbest* values. The values corresponding to C_1 and C_2 are set to 2 as recommended by Kennedy and Eberhart (1995). The reason for setting $C_1 = C_2 = 2$ is the social learning and cognition learning have the same effect in velocity update.

4.4 Modified self-adaptive inertial weight with convergence logic

Since each of the particles associated with the swarm has inherited variant differences in each state of iteration, it is logical to have self-adaptive inertial weight at the individual level. In accordance with the particle behavior of mechanics, the velocity update is expressed in Eqs. 6 and 7 as:

$$V^j_{i,n+1} = wV^j_{i,n} + F^j_{i,n}, \tag{6}$$

where

$$F^j_{i,n} = C_1 r^j_{i,n}\left(P^j_{i,n} - X^j_{i,n}\right) + C_2 r^j_{i,n}\left(G^j_{i,n} - X^j_{i,n}\right). \tag{7}$$

The value of $F^j_{i,n}$ represents the acceleration of the particle moving toward the region of $P^j_{i,n}$ and G^j_n in the given search space. If the regions of $V^j_{i,n}$ and $F^j_{i,n}$ point in different directions, then the particle is said to lie far away from the optimal region of $P^j_{i,n}$ and G^j_n, respectively. As a result, the particle is said to modify its jth component for velocity and the corresponding inertial weight $w^j_{i,n}$ should be set to a smaller value. In another way, smaller values of $V^j_{i,n}$ and $F^j_{i,n}$ imply that it is not crucial to modify the particle velocity and the value of $w^j_{i,n}$ can be set to a larger value. The acceleration of the particle in accordance with the convergence speed and the fixation of inertial weight follow the following conditions:

- If the values of $V^j_{i,n}$ and $F^j_{i,n}$ are large, then the jth component is in correct direction, and it is said to speed up the inertial weight which is then set to a larger value of index.
- If the values of $V^j_{i,n}$ and $F^j_{i,n}$ are small, then the jth component of the particle is in near-optimal position, and the weight is set to a smaller value of index.

Upon these conditions, the convergence of the particle is assigned in accordance with the particle's position and its movement in velocity. Therefore, the value of $w^j_{i,n}$ is considered to be the function of $\left|F^j_{i,n}\right|$ and $\left|V^j_{i,n}\right|$. The incorporation of the inertial weight factor is depicted in Eq. 6 and the value of $F^j_{i,n}$ in Eq. 7. The value of inertial weight can be chosen accordingly with the linear or nonlinear function of the number of iterations. Table 4 describes the various stages of inertial weights with regard to the number of iterations and accuracy.

It has been observed that with inertial weight for value 0.9, the estimated accuracy is found to 98.60% with external cal-

Table 3 Attribute Description

S. no	Attribute name	Description
1.	Age	Age of the person
2.	Fasting plasma glucose (FPG)	Taken before the consumption of fluids or breakfast
3.	Postprandial plasma glucose(PPG)	Taken 1½ h after the consumption of diet
4.	Glycosylated hemoglobin (A1c)	To identify average plasma glucose concentration for 3 months in the human body
5.	Mean blood glucose (MBG)	Mean glucose level for 3 months in mg/dl
6.	Total cholesterol	Below 200 mg/dl would be considered as normal
7.	Triglyceride level (TGL)	A type of fat which has been determined by food consumption
8.	Low-density lipoprotein cholesterol (LDL)	Bad cholesterol deposited over arterial blood vessels
9.	High-density lipoprotein (HDL) cholesterol	Good cholesterol which maintains the arteries at an optimal level
10.	Very low-density lipoprotein (VLDL) cholesterol	VLDL is one-fifth of the triglyceride level, although this is less accurate if the triglyceride level is greater than 400 mg/dl.
11.	Non-high-density lipoprotein (NHDL) cholesterol	NHDL cholesterol is the total cholesterol minus HDL cholesterol NHDL cholesterol = total cholesterol − HDL cholesterol
12.	Blood urea (BU)	Blood urea (BU) test is used to determine how well kidneys are working
13.	Streptokinase (SK)	It is an enzyme test (Streptococci)
14.	Albumin/creatinine ratio (ACR)	The urine albumin test or ACR is used to screen people with diabetes, chronic disorder and hypertension
15.	Total protein	Total protein which comprises of albumin and globulin
16.	Albumin	It is a enzyme test (produced in the liver)
17.	Globulin	An enzyme which is responsible for the effective functioning of circulatory system
18.	Serum glutamic oxaloacetic transaminase (SGOT)	An enzyme which is present in liver and heart cells
19.	Serum glutamic–pyruvic transaminase (SGPT)	An enzyme which is present in liver and heart cells
20.	Alkaline phosphatase (ALP) test	A protein test in blood level found in all tissues
21.	Glutamyl transpeptidase (GGT)	The GGT value measure corresponds to the level of enzyme GGt in blood
22.	Haemo (HB)	The protein red blood cells which carry oxygen from lungs to tissues and bring back carbon dioxide to lungs
23.	Class label	The label value which depicts the occurrence of the disease

Table 4 Selection of inertial weight with improvement in accuracy

S. no.	Iteration	Inertial weight values	Accuracy
1.	1	0.4	76.54
2.	2	0.5	81.86
3.	3	0.6	87.32
4.	4	0.7	92.21
5.	5	0.8	96.84
6.	**6**	**0.9**	**98.60**
7.	7	1.0	97.24
8.	8	1.1	93.00
9.	9	1.2	90.84
10.	10	1.3	89.96

Bold denotes the maximum accuracy value obtained with respect to the inertial value of about 0.9

ibrated iteration value of 6, which is highlighted in Table 3. The detailed internal estimates for the iteration value six are given in Table 4 with all the observed parametric measures. The highest accuracy value, kappa, WMR, WMP, correlation with minimum classification error, absolute error and RMSe are obtained for the specific particle number, which are highlighted in Table 5. Literature experiences for the choice of inertial weight have also been analyzed over the benchmark data as implemented by Shi and Eberhart (1998a). For the benchmark cases, the values for the inertial weight have been varied below 0.8 and above 1.2. The exploration ability of the algorithm is found to be worse than that of the original PSO algorithm. The same has been reflected in our real-world diabetic data as depicted in Fig. 4.

When the values of w are fixed between 0.8 and 1.2, the algorithm has a better way for search space exploration to fix the global optimum solution with a minimum number of iterations. The experimental work by Shi and Eberhart (1998b) also states that the range for inertial weight values has given a significant improvement in the performance of PSO algorithm in a decreased level of time with iterations.

Table 5 Observed measures for performance metrics

S. no.	Particle number	Accuracy (%)	Classification error (%)	Kappa	WMR (%)	WMP (%)	Absolute error	RMS error	Correlation
1	0	91.40	8.60	0.809	90.12	91.36	0.106	0.266	0.814
2	1	91.00	9.00	0.803	89.44	91.19	0.115	0.265	0.806
3	2	92.00	8.00	0.824	90.95	91.77	0.102	0.255	0.827
4	3	96.40	3.60	0.922	96.34	95.93	0.037	0.157	0.923
5	4	89.80	10.20	0.771	87.29	90.51	0.146	0.3	0.777
6	5	96.40	3.60	0.922	96.34	96.01	0.042	0.165	0.923
7	6	91.00	9.00	0.804	90.17	90.44	0.111	0.283	0.806
8	7	91.00	9.00	0.805	90.17	90.63	0.109	0.281	0.808
9	8	94.80	5.20	0.886	94.11	94.68	0.066	0.199	0.888
10	9	95.60	4.40	0.903	94.98	95.61	0.052	0.174	0.906
11	10	91.00	9.00	0.801	89.32	91.31	0.114	0.275	0.806
12	11	95.20	4.80	0.897	95.28	94.89	0.049	0.168	0.901
13	12	91.80	8.20	0.816	89.70	92.83	0.117	0.260	0.824
14	13	93.20	6.80	0.853	92.5	93.19	0.079	0.221	0.857
15	14	95.80	4.20	0.909	95.62	95.32	0.049	0.175	0.909
16	15	92.60	7.40	0.837	91.42	92.75	0.103	0.240	0.841
17	16	89.20	10.80	0.760	87.19	89.63	0.150	0.310	0.767
18	**17**	**98.60**	**1.40**	**0.970**	**98.66**	**98.37**	**0.015**	**0.071**	**0.970**
19	18	90.00	10.00	0.776	87.69	90.54	0.137	0.295	0.782
20	19	88.80	11.20	0.756	87.73	88.06	0.142	0.315	0.758

The level of decreasing factor concerning time for inertial weight is given in Eq. 8 as:

$$w_n = \frac{(w_{\text{initial}} - w_{\text{final}})(n\text{max} - n)}{n\text{max} + w_{final}}, \tag{8}$$

wherew_n—the value corresponding to inertial weight at nth iterationw_{initial}—the initial value of inertial weightw_{final}—the final value of inertial weight$n$—iteration number$n\text{max}$—maximum iteration number.

Massive forms of experimental investigations have been made by Shi and Eberhart (1999) in accordance with decreasing variations of inertial weight. The convergence rate of PSO algorithm is not sensitive toward the number of iterations and the population size. Also, it is to be importantly noted that PSO algorithm may lack global search capability if the value of the chosen inertial weight is small. Therefore, as a consequence, PSO algorithm can be solved with the dynamic forms of real-world applications. $Rand_1$ and $Rand_2$ are the particle generation random numbers which are set to the value around $[0 - 1]$, respectively. During iteration, the particle's best position is determined by the value of *pbest* and the best value obtained so far by some particles is *gbest* for the given set of the population as illustrated by Ebrahim Sorkhabi et al. (2016). Once the features are selected, accuracy is determined using the classifier by formulating test and training dataset. The validation of the model is set to tenfold for the mechanism of cross-validation. Hence, the one part of the entire data is used up for testing phase, and remaining nine parts are used for the training phase. The execution continues until all the segments of the tuples are evaluated for cross-validation.

4.5 Metrics for evaluation

The performance of the method is evaluated through the following metrics:

1. Accuracy
   If we use the entire set of training data to model the classifier performance, then the observed result would be optimistic in nature. For our problem, we validate the model with the test data tuples. These test tuples are being randomly selected which are independent of the training data tuples. Meanwhile, these test tuples are not for modeling the classifier; they are used to estimate the classifier performance. The accuracy of a classifier is defined as the percentage of test data tuples that have been correctly classified by the classifier. The label concerned with each of the tuples in a test data record is compared with that of the learned classifier class prediction for the tuple as suggested by Baldi et al. (2000). The evaluation of accuracy is estimated in Eq. 9 as follows:

$$\text{Accuracy} = \left(\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}\right). \tag{9}$$

2. Classification error
   The classification error in a classifier model is defined as the total number of incorrectly classified instances or tuples of record during evaluation. The total number of misclassified instances defined the error rate of the developed model in Eq. 10 as:

$$E_i = \left(\frac{n}{N}\right), \tag{10}$$

   where n—total number of tuples incorrectly classified and N—total number of tuples in the given dataset of record.
3. Kappa statistics.
   It is a measure of the degree of non-random agreement between observers or measurements of the same categorical variable. The calculation is given in Eq. 11 as:

$$K = \left(\frac{p(A) - p(E)}{1 - p(E)}\right). \tag{11}$$

4. Weighted mean recall (WMR) and weighted mean precision (WMP)
   F-score is a statistical measure of test accuracy which considers both precision p and recall r to determine the score value. The value p determines the fraction of retrieved instances that are relevant, and r determines the fraction of relevant instances that are retrieved. The measure is given in Eqs. 12 and 13 as:

$$\text{precision} = \left(\frac{\text{TP}}{\text{TP} + \text{FP}}\right), \tag{12}$$

$$\text{recall} = \left(\frac{\text{TP}}{\text{TP} + \text{FN}}\right). \tag{13}$$

   The F-score is then calculated with the weighted average of precision and recall.

$$F\text{score} = 2 \cdot \left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right). \tag{14}$$

5. Absolute error
   It is measured as the difference between the inferred values for a tuple of record to that of the actual value. It is given in Eq. 15 as:

$$\Delta t = t_0 - t. \tag{15}$$

6. Root mean squared error (RMS error)
   It is a statistical measure of the size of a varying amount likewise called as the quadratic mean. It can be computed for a progression of discrete values or a persistently varying function. It is given in Eq. 16 as:

$$t_{\text{RMS}} = \sqrt{\frac{\sum_{i=1}^{n} t_i^2}{n}}. \tag{16}$$

   For a variant of t over a continuous distribution $P(x)$, RMS is given in Eq. 17 as:

$$t_{\text{RMS}} = \sqrt{\frac{\int [p(t)]^2 \mathrm{d}t}{\int [p(t)] \mathrm{d}t}}. \tag{17}$$

4.6 Cross-validation and results obtained

Cross-validation deals with the partition in the initial set of data into k mutual subset. If D is the dataset that deals with the process of data classification, it is segregated into the number of folds such as $D_1, D_2, D_3 \ldots D_k$. The training and testing phases for the entire dataset are performed k number of times. In the first iteration, the i partition D_i is meant for the test dataset, and the remaining separations are used up to train the data model. Hence, during the first iteration in developing the model the partition $D_1, D_2, D_3 \ldots D_k$ serves to be the training datasets to derive the first data model. During the second iteration the partition other than D_2will be the training data with test on D_2. The accuracy of the method is then determined in accordance with the total number of correctly classified records.

The dataset used consists of 732 instances of records and 23 attributes involved in the evaluation and development of the model. Out of 23 attributes, postprandial plasma glucose (PPG), glycosylated hemoglobin (A1c), mean blood glucose (MBG) and fasting plasma glucose (FPG) are the attributes selected for evaluation. By setting up tenfold cross-validation, over the number of iterations and generations, the tree obtained by the classifier model is depicted in Fig. 5. The observance of the results with the metric value is described in Table 4. The generated tree provides the root node to be postprandial plasma glucose (PPG), with a split value of about > 269 and ≤ 269.500 made by glycosylated hemoglobin (A1c) and mean blood glucose (MBG). The split is made with respect to binary segregates of the data tuples. Data discretization among the numeric attribute values involve data partitioning with break points. It represents the split among the class labels with a majority in one side and the remaining part in another. Each set of split value is then determined by the break points with same or different class label values.

There are two major cases to be considered for the split value and break point determination:

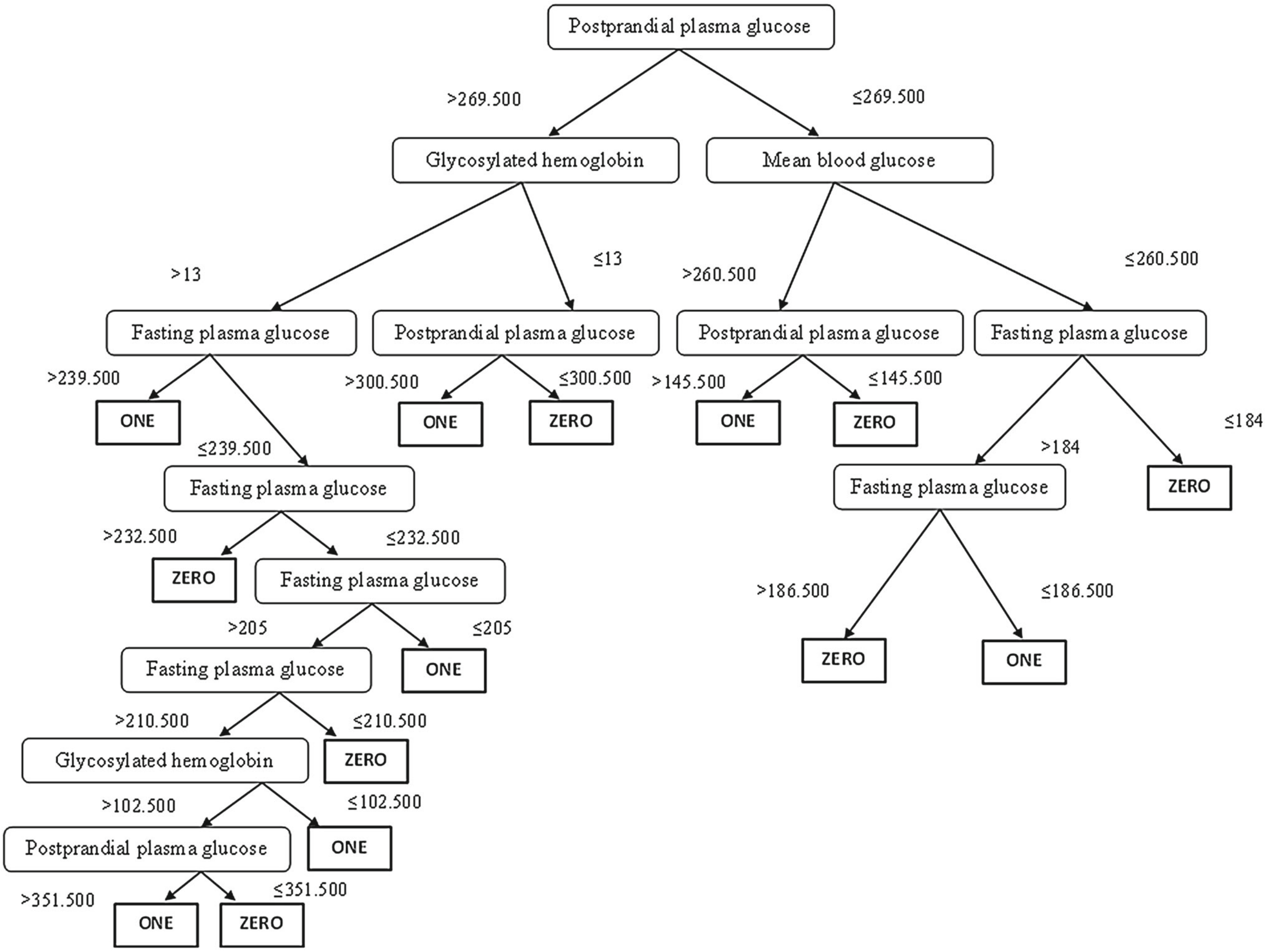


Fig. 5 Generated tree for classification of selected features

- During data partition, if the preceding class values are of the same type, then the partition can be merged with that class value.
- During data partition, if adjacent consists of the same sort of majority of the similar class label, then they can be merged without breaking the rule. The tree generation process continues until all the branches have been attained with a leaf node, i.e., labeled.

5 Formulating the association among the risk factors by mathematical model using Fisher's linear discriminant analysis

5.1 Association of type II diabetic risk with attribute measurements

The determined level of accuracy from the set of iterations has been got up to 98.60% with the attributes such as postprandial plasma glucose (PPG), glycosylated hemoglobin (A1c), mean blood glucose (MBG) and fasting plasma glucose (FPG). The association and correlation among the attributes can be determined by using Karl Pearson's correlation coefficient. If the change in one variable affects the change in another variable, then the variables are said to be correlated.

The correlation between the two random variables X and Y which is denoted as $r(X, Y)$, and it is defined in Eq. 18 as:

$$r(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{18}$$

For the given set of data values (x_i, y_i), where $i = 1, 2, 3 \ldots, n$,

then,

$$Cov(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}], \tag{19}$$

$$\sigma_{X^2} = \frac{1}{n} \sum (x_i - \overline{x})^2, \tag{20}$$

$$\sigma_{Y^2} = \frac{1}{n} \sum (y_i - \overline{y})^2. \tag{21}$$

Table 6 Determination of correlation among attributes

Variables	MBG	A1c	FPG	PPG
MBG	**1.000**	0.719	0.732	0.733
A1c	0.719	**1.000**	0.512	0.507
FPG	0.732	0.512	**1.000**	0.737
PPG	0.733	0.507	0.737	**1.000**

Bold denotes the maximum value of correlation

The limits of correlation among the determined attributes PPG, MBG, FPG and A1c are given by Eq. 22 as:

$$r(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{\frac{1}{n}\sum(x-\overline{x})(y-\overline{y})}{\left[\frac{1}{n}\sum(x-\overline{x})^2\cdot\frac{1}{n}(y-\overline{y})^2\right]^{1/2}}. \tag{22}$$

The set of correlation has been found among (MBG, A1c), (MBG, FPG), (MBG, PPG) and (FPG, PPG) as illustrated in Table 6. Hence, mean blood glucose has its strong correlation toward A1c, FPG, and PPG. Similarly, postprandial plasma glucose has its strong correlation toward MBG and FPG, respectively. With the set of possible combinations, the following combinations among the attribute seem to have a correlated value with its varying discretions. The observed R^2 values among the correlated attributes are 0.5392, 0.5058, 0.4922 and 0.1658, respectively. The comparison among the accuracies with the proposed model PSO–J48 is depicted in Fig. 6. In Figs. 7, 8, 9 and 10, exponential curves should meet all the points that lie on it. Meanwhile, curve fitting provides the process of capturing the trend in the data by assigning a single function across the entire range. The motto behind curve fitting is to determine the coefficients ‘a’ and ‘b’ such that the function fits with the data in an affordable

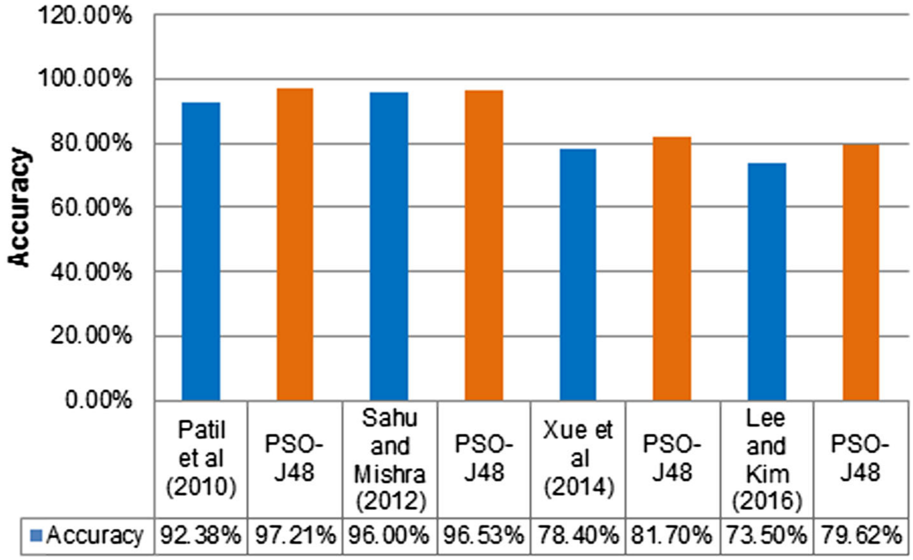


Fig. 6 Comparisons among accuracies with existing approaches against the proposed model

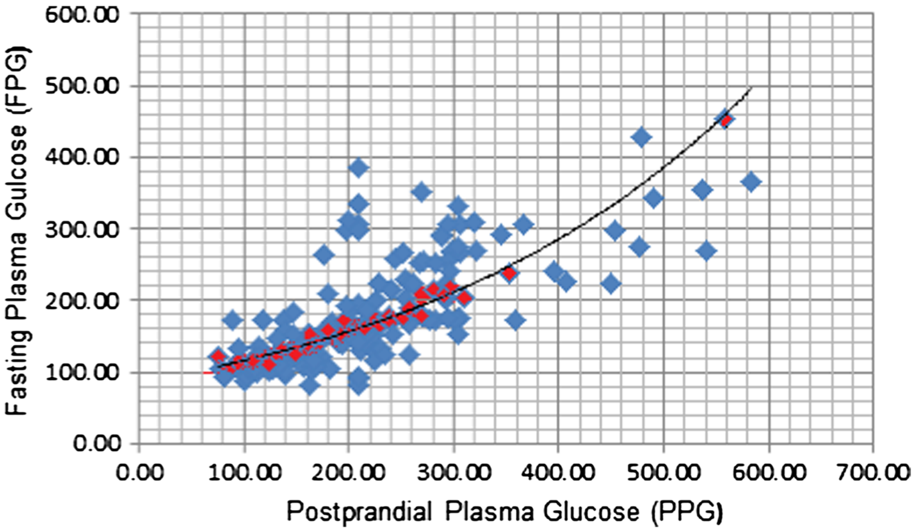


Fig. 7 Observed exponential curve between PPG and FPG with $y = 85.918e0.003x$ and $R^2 = 0.5392$

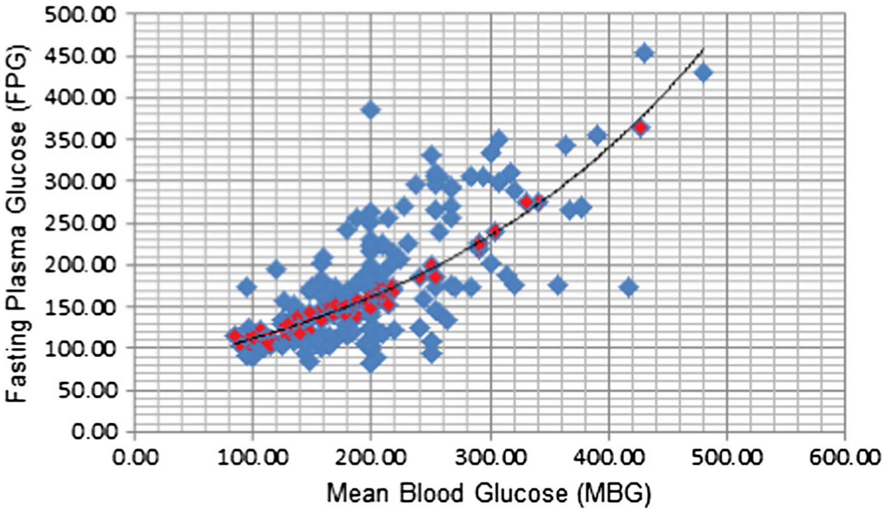


Fig. 8 Observed exponential curve between MBG and FPG with $y = 76.984e0.0037x$ and $R^2 = 0.5058$

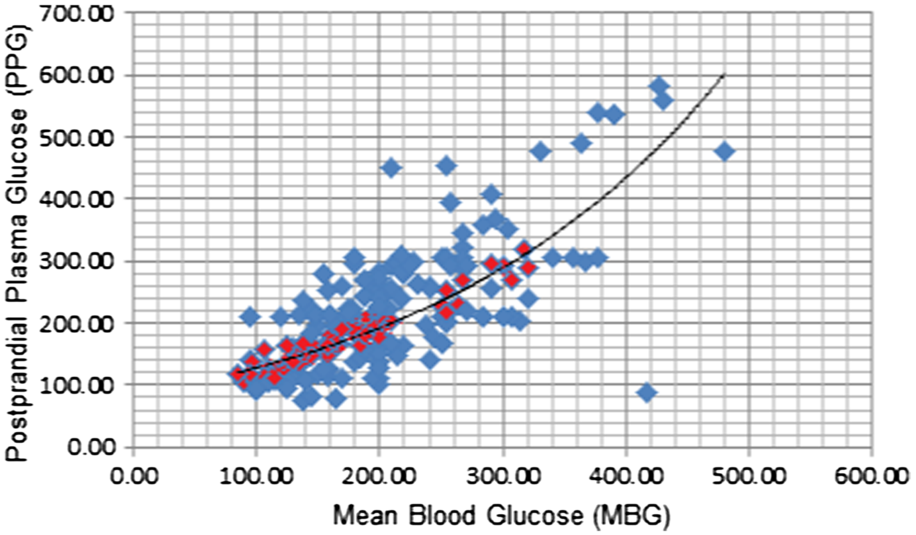


Fig. 9 Observed exponential curve between MBG and PPG with y = 85.281e0.0041x and $R^2 = 0.4927$

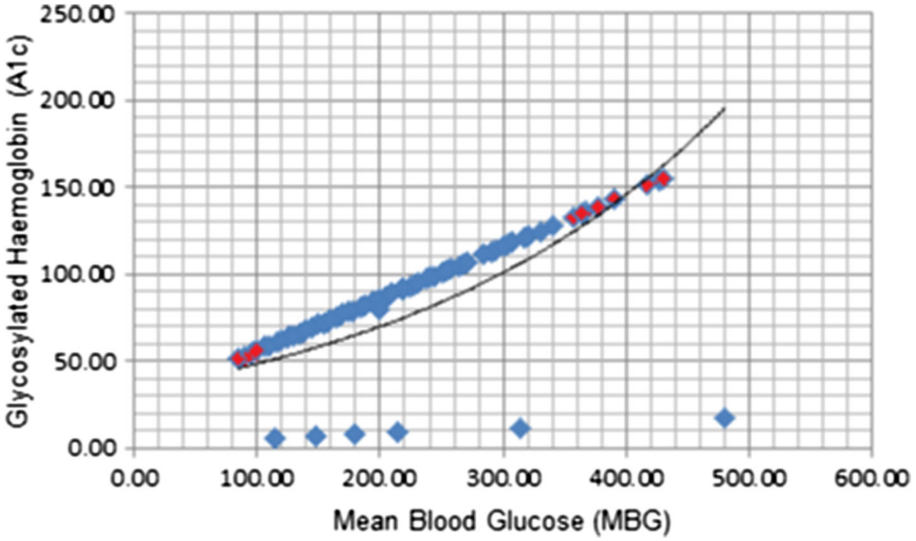


Fig. 10 Observed exponential curve between MBG and PPG with $y = 33.609e0.0037x$ and $R^2 = 0.1658$

way. The R^2 computation determines the percentage of the response variable and its variation by an exponential model. The more variance encountered by the model then closer is the data points that tend to fit the data model.

Table 7 Comparison of the proposed model with the existing methods for various datasets

S. no.	Method	Dataset	Total no. of features	No. of features selected	Accuracy (%)
1.	Xue et al. (2014) [48]	Lung UCI dataset	56	6	78.40
	Proposed model PSO-J48			5	81.70
2.	Patil et al. (2010) [31]	Pima Indian diabetic dataset	5	3	92.38
	Proposed model PSO-J48			3	97.21
3.	Lee and Kim (2016) [26]	Type II diabetic dataset	15	3	73.50
	Proposed model PSO-J48			3	79.62
4.	Sahu and Mishra (2012) [34]	Benchmark colon cancer dataset	2000	5	96.00
	Proposed model PSO-J48			4	96.53

5.2 Comparison of the proposed model with existing approaches

The results obtained using proposed PSO–J48 has been evaluated against various approaches given by authors corresponding to the work of domain. The work has been tested against the dataset used by authors for other such models that they have developed. Table 7 provides the experimental results observed in accordance with the various other datasets in comparison with their approaches. Figure 6 explains in accordance with the accuracy of the existing approaches against the proposed model.

Table 8 Measured pooled within-class covariance matrix

	MBG	A1c	FPG	PPG
MBG	3592.459	984.691	2091.529	2694.772
A1c	984.691	668.195	524.821	662.446
FPG	2091.529	524.821	3397.268	2605.972
PPG	2694.772	662.446	2605.972	5604.501

5.3 Formulating the association among the risk factors by mathematical model using Fisher's linear discriminant analysis

Linear discriminant analysis or Fisher's linear discriminant relates to the classification problem, where two or more groups of samples are grouped into clusters of records in which the newer set of observation falls into the known population based upon signified characteristics (Fisher 1936). The principal component analysis makes over the investigation in a lower-dimensional space. Hence, information with maximum variances is not suitable for classification problem. The discrimination among the data into a different set of classes will not be made by principal component analysis, and therefore, the linear discriminant analysis has been chosen for data evaluation (Fisher 1938).

Consider the set of d-dimensional samples of $x_1, x_2, x_3 \cdots x_n$, let n_1, n_2 correspond to the subset of records from D_1, D_2 with ω_1 and ω_2 as labels, respectively. The linear combination of components corresponding to x for y is given as $y = w^t x$. With the selected set of features, the formulation among the best separation of classes for MBG, PPG, FPG and A1c has to be determined. The separation of classes for the attributes has to be suited in the best direction corresponding to w (Cristianini 2004). The mean for 732 subjects in its dimensional space is given in Eq. 23 as:

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x. \tag{23}$$

For the set of subjects, the sample mean of projected points is given by:

$$\overline{m_i} = \frac{1}{n_i} \sum_{y \in y_i} y = \frac{1}{n_i} \sum_{x \in D_i} w^t x = w^t m_i. \tag{24}$$

The separation between the anticipated methods for the two classes zero and one is ascertained by utilizing the following equation:

$$|\overline{m_1} - \overline{m_2}| = \left|w^t (m_1 - m_2)\right|. \tag{25}$$

The defined scatter for the projected variance is for the pooled data, which is given as: $\left(\frac{1}{n}\left(\overline{s_1}^2 + \overline{s_2^2}\right)\right)$.

Table 8 defines the projected pooled data value for the best separation of records among the classes.

Table 9 Measured within-class covariance matrix for class 0

	MBG	A1c	FPG	PPG
MBG	2434.727	722.569	923.531	1330.803
A1c	722.569	546.624	175.152	283.505
FPG	923.531	175.152	1914.389	1544.411
PPG	1330.803	283.505	1544.411	2984.082

Table 10 Measured within-class covariance matrix for class 1

	MBG	A1c	FPG	PPG
MBG	5655.680	1451.827	4173.043	5125.532
A1c	1451.827	884.850	1147.975	1337.765
FPG	4173.043	1147.975	6039.942	4497.804
PPG	5125.532	1337.765	4497.804	10,274.409

The following step is to locate the linear function $w^t x$ in which

$$J(w) = \frac{|\overline{m_1} - \overline{m_2}|^2}{\left(\overline{s_1}^2 + \overline{s_2^2}\right)}. \tag{26}$$

This ought to be considered to be higher and independent of $\|w\|$. The within-class scatter matrix can be determined using $S_w = S_1 + S_2$, and between-class matrix is found using $S_B = (m_1 - m_2)(m_1 - m_2)^t$. The observed class covariance matrix is given in Tables 9 and 10.

The within-class scatter matrix is proportional to the sample covariance for the pooled dimensional data. Regarding S_B and for S_W, the function is given as $J(w) = \frac{w^t S_B w}{w^t S_w w}$. Hence, the vector w that maximizes to formulate the condition should satisfy $S_B w = \lambda S_W w$, which implies to the formulation of $S_W^{-1} S_B w = \lambda w$, which again denotes to eigenequation. Hence, we have obtained the value of Fisher's discriminant analysis which provides the ratio of between-class scatter to within-class scatter and proves to be different as illustrated by Flury and Riedwyl (1988).

5.3.1 Test interpretation

The esteem of elucidation can be decided by the hypothesis test. Let H_0 be the within-class covariance matrix to be assumed as in equal state and H_1 be the within-class covariance matrix in a different state.

H_0—null hypothesis
H_1—alternate hypothesis

From the results, it is observed that the p value is lower than that of the level of alpha which is 0.05. So we can reject the null hypothesis H_0 and accept the alternate hypothesis H_1. Therefore, it has been concluded from the mathematical model that the within-class covariance matrices are different. The state of risk is to reject the null hypothesis, while it is true of lower than 0.01%. Table 11 provides the results observed with Fisher's discriminant analysis with observed and critical values.

Table 11 Fisher's F asymptotic approximation

F (observed value)	22.519
F (critical value)	65,326.551
DF1	10
DF2	651,387
p value	<0.0001
Alpha	0.05

The goodness of fit computed over χ^2 analysis among the observed and the expected frequencies of (A_i, B_j) is computed using Eq. 27 as:

$$\chi^2 = \sum_{i=1}^{C}\sum_{j=1}^{R} \left(O_{ij} - E_{ij}\right)^2 \Big/ E_{ij}. \tag{27}$$

The χ^2 statistic tests the hypothesis that H_0 and H_1 are independent. The test is based upon significance level of $(r - 1) * (c - 1)$ degrees of freedom. The computed degrees of freedom for the statistical test are 10, and for 10 degrees of freedom, the probability is observed to be 18.30 with the observed $\chi^2 = 225.19$, respectively. Since our value is more prominent than that of the probabilistic value $p < 0.0001$, we can reject the null hypothesis and conclude that the attributes PPG, MBG, FPG and A1c are dependent on each other.

5.3.2 Test evaluation

The test evaluation has been done using receiver operating curve (ROC) curve which provides the medium for sensitivity and specificity report. The points in the curve determine the set of sensitivity/specificity pairs which corresponds to the decision threshold value. The area provided by the curve provides the mechanism of distinguishing among the signified class labels which reveal diseased and normal conditions. The discrimination is determined for every possible cutoff criterion in which some of the cases with the disease are to be correctly classified (true positive), and some of the cases are correctly classified for those without the disease (true negative). But, there will be cases in which cases with the disease will be wrongly classified as negative and without the disease will be classified as positive. Table 12 provides the best separation for the confusion matrix in classifying the data for the population with the disease and without the disease.

Table 12 Testing criteria

Test criteria	Disease status	
Positive cases	True positive (TP)	False positive (FP)
Negative cases	False negative (FN)	True negative (TN)

Table 13 ROC analysis report

Area under the curve (AUC)	0.887
Standard error	0.0281
Sensitivity	98.66
Significance level p (area = 0.5)	<0.0001

The statistics corresponding to sensitivity and specificity is given in Eqs. 28 and 29 as:

$$\text{positive\_likelihood\_ratio} = \frac{\text{Sensitivity}}{1 - \text{specificity}}, \tag{28}$$

$$\text{negative\_likelihood\_ratio} = \frac{1 - \text{sensitivity}}{\text{Specificity}}. \tag{29}$$

Sensitivity defines the state as the estimate of the probability that signifies the test result as positive if the disease is present (percentage of TP). Specificity defines the state as negative if the probability that signifies the test result is not present (percentage of TN).

5.3.3 ROC analysis

The efficacy of the method has been tested for ROC analysis in which the cumulative distribution function signifies the value of ∞ to the specified threshold value. The curve provides the estimate in accordance with probability detection in the y-axis and the cumulative distribution function in the x-axis. The class distribution in accordance with the binary classification problem has been evaluated for analysis. Meanwhile, the possible way of determining optimal modes can be taken into consideration for the experimental analysis as given by Greiner et al. (2000). The ROC curve in Fig. 10 signifies the plot for sensitivity and specificity for different cutoff points as illustrated by DeLong et al. (1988). A decision threshold value is used by each point which represents the pair of data. The perfect test interpretation and discrimination have been interpreted if there is no form of overlap among the distributions. Table 13 provides the results obtained using ROC.

Hence, from the observed results, the AUC equals to 0.887 which is nearer to the value 1. Therefore, among the two sets of distribution there occur no overlaps and the values of sensitivity and specificity pairs in which the ROC of the curve lies at the upper left corner corresponding to the plot as in Fig. 11. Therefore, the distribution concerning the cumulative function and the probability estimate values indicate that there exist no controversies among the class distribution. The probability of detection is also near to the optimal value of 1. Therefore, for the considered dataset, the perceived model proves well to be measured regarding accuracy, sensitivity and specificity also with the model tested by Gardner and Greiner (2006).

The proposed model using PSO–J48 algorithm provides an insight into the observed data which corresponds to the region of Theni. Nature, habits, locality, likelihood and dietary conditions of the people may vary from region to region. Each group of people may adhere to policies and framework in which the obsolete way of living can vary from time to time. The objective of this research work is to develop a predictive model for type II diabetic risk prediction using PSO and decision tree algorithm. The developed model best spots the risk factors that are more predominant and that are having strong correlation with the observed data. Medical practitioners can well utilize the developed method to observe the predominant risk factor that corresponds to the specified disease. Also, the model can well be adhered to in practice for the determination of correlation patterns that exist between the risk factors and its test case criterions concerning accuracy, sensitivity, specificity and its p value as depicted by Smith and Slenning (2000).

From the analysis report, it has been observed that the attributes MBG, PPG, FPG and A1c have a resilient dependency among them regarding test measures, test interpretation and evaluation.

6 Conclusion and future work

Data classification and optimization in prediction play a significant role in medical data analysis. Collecting and processing medical data with its range values seems to be a challenging problem. The entire formulation of the research work signifies the development of a decision support model with an improved PSO and decision tree algorithm. The improvement is made in fixing the operating parameter, i.e., modified self-adaptive inertial weight with convergence logic for the particles to explore in the given search space to find a near-optimal solution. The dataset with 732 records having 23 attributes including the class label has been used for analysis. The test result demonstrates that the proposed method produces four important risk factors with improved accuracy of about 98.60%. The effectiveness of the proposed model has been validated against the data and the methodology given by various authors. Meanwhile, the association among the selected features such as MBG, PPG, A1c and FPG has been determined. A strong correlation has been observed for MBG with A1c, FPG and PPG, and also PPG has its strong correlation toward MBG and FPG, respectively.

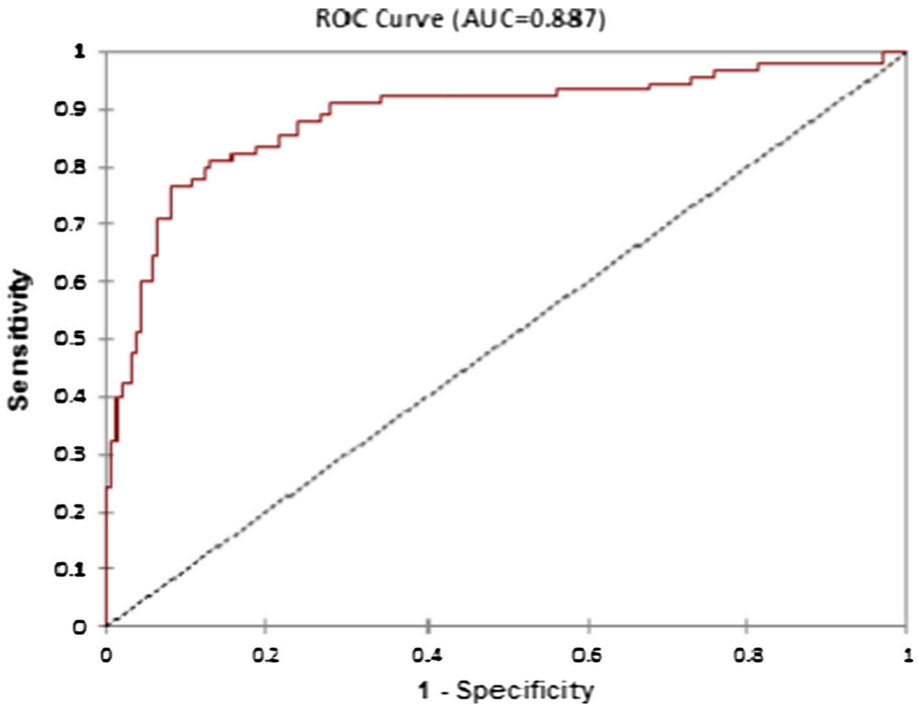


Fig. 11 ROC for sensitivity versus specificity

The observed R^2 values among the correlated attributes are 0.5392, 0.5058, 0.4927 and 0.1658 which are also found to be significant.

A mathematical model has been developed using Fisher's LDA for the discovered attributes. The within-class covariance matrix for class zero and one has been calculated, and the test interpretation results prove that the value $\chi^2 = 222.19$ which is significantly greater than the probabilistically observed value $p < 0.0001$. Therefore, the alternate hypothesis is accepted for the signified F value across the probabilistic measure in degrees of freedom. The test interpretation with ROC analysis value has been observed at 0.887 with significant level <0.0001. Hence, for the data corresponding to type II diabetes of 732 subjects the risk prevalence has been found across MBG, PPG, A1c and FPG. The developed model can be used by physicians for earlier determination of the disease and to reduce clinical test which, in addition, reduces the cost of treatment.

The future work can focus on the different combination of feature selection algorithms and data classification techniques. The improvement can be formulated with the adoption of the different combinations of parametric values. Data optimization techniques such as ant colony optimization, genetic algorithm and swallow optimization can be adopted. Therefore, the target measure in the determination of the risk factors in accordance with the concerned disease can be known in advance, and the implication of the disease can be prohibited. The nature of location, likelihood and dietary habits of the patients vary significantly from region to region. Similar to this research work, region-based analysis can be performed in accordance with severity of communicable and non-communicable diseases.

Acknowledgements We would like to sincerely and heartily thank Dr. M. Venkatesh, M.B.B.S, M.D. (Gen Medicine), Senior Assistant Professor, Department of General Medicine, Government Theni Medical College, Theni. He has supported the entire research work on the risk factor prediction in type II diabetes with a measure of relevance feedback at real-time analysis. He handled various projects in the Department of General Medicine and is fruitfully working toward the technological impacts of the disease-specific syndromes.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Abdel-Kader RF (2010) Genetically improved PSO algorithm for efficient data clustering. In: 2010 second international conference on machine learning and computing, IEEE. http://dx.doi.org/10.1109/icmlc.2010.19

Anon (2016) New economic reality: the rise of big data and big analytics. Virtual Compet. https://doi.org/10.4159/9780674973336-002

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Chen KH et al (2014) Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinform 15(1):49. https://doi.org/10.1186/1471-2105-15-49

Chuang L-Y, Tsai S-W, Yang C-H (2011) Improved binary particle swarm optimization using catfish effect for feature selection. Expert Syst Appl 38(10):12699–12707. https://doi.org/10.1016/j.eswa.2011.04.057

Collen MF (1994) The origins of informatics. J Am Med Inform Assoc 1(2):91–107. https://doi.org/10.1136/jamia.1994.95236152

Cristianini N (2004) Fisher discriminant analysis (linear discriminant analysis). Dictionary of bioinformatics and computational biology. Wiley. http://dx.doi.org/10.1002/9780471650126.dob0238.pub2

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44(3):837. https://doi.org/10.2307/2531595

Eberhart RC, Shi Y (2001) Tracking and optimizing dynamic systems with particle swarms. In: Proceedings of the 2001 congress on evolutionary computation (IEEE Cat No01TH8546), IEEE. http://dx.doi.org/10.1109/cec.2001.934376

Ebrahim Sorkhabi A, Deljavan Amiri M, Khanteymoori AR (2016) Duality evolution: an efficient approach to constraint handling in multi-objective particle swarm optimization. Soft Comput 21(24):7251–7267. https://doi.org/10.1007/s00500-016-2422-5

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fisher RA (1938) The statistical utilization of multiple measurements. Ann Eugen 8(4):376–386. https://doi.org/10.1111/j.1469-1809.1938.tb02189.x

Flury B, Riedwyl H (1988) Linear discriminant analysis for two groups. In: Multivariate statistics. Springer, Netharlands. https://doi.org/10.1007/978-94-009-1217-5_7

García-Nieto J, Alba E (2010) Restart particle swarm optimization with velocity modulation: a scalability test. Soft Comput 15(11):2221–2232. https://doi.org/10.1007/s00500-010-0648-1

Gardner IA, Greiner M (2006) Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. Vet Clin Pathol 35(1):8–17. https://doi.org/10.1111/j.1939-165x.2006.tb00082.x

Gendreau M, Potvin JY (2010) Handbook of metaheuristics. International series in operations research and management science. Springer, New York. https://doi.org/10.1007/978-1-4419-1665-5

Greiner M, Pfeiffer D, Smith R (2000) Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med 45(1–2):23–41. https://doi.org/10.1016/s0167-5877(00)00115-x

Gupta R, Misra A (2007) Review: type 2 diabetes in India: regional disparities. Br J Diabetes Vasc Dis. 7(1):12–16. https://doi.org/10.1177/14746514070070010301

Han J (1996) Data mining techniques. ACM SIGMOD Record. Association for Computing Machinery (ACM); 25(2):545. http://dx.doi.org/10.1145/235968.280351

Hapsara HR (2005) World Health Organization (WHO): global health situation. In: Encyclopedia of biostatistics. https://dx.doi.org/10.1002/0470011815.b2a17156

Ievers-Landis CE, Walders-Abramson N, Amodei N, Drews KL, Kaplan J, Levitt Katz LE et al (2015) Longitudinal correlates of health risk behaviors in children and adolescents with type 2 diabetes. J Pediatr 166(5):1258–1264. https://doi.org/10.1016/j.jpeds.2015.01.019

Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS (2010) Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans Inf Technol Biomed 14(3):559–566. https://doi.org/10.1109/titb.2009.2038906

Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: Proceedings of ICNN'95—international conference on neural networks, IEEE. http://dx.doi.org/10.1109/icnn.1995.488968

Khan HMR, Mende S, Rafiq A, Gabbidon K, Reddy PH (2017) Methods needed to measure predictive accuracy: a study of diabetic patients. Biochim Biophys Acta (BBA) Mol Basis Dis 1863(5):1046–1053. https://doi.org/10.1016/j.bbadis.2017.01.007

Lee BJ, Kim JY (2016) Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J Biomed Health Inform 20(1):39–46. https://doi.org/10.1109/jbhi.2015.2396520

Lin S-W, Ying K-C, Chen S-C, Lee Z-J (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Syst Appl 35(4):1817–1824. https://doi.org/10.1016/j.eswa.2007.08.088

Liu Y, Wang G, Chen H, Dong H, Zhu X, Wang S (2011) An improved particle swarm optimization for feature selection. J Bionic Eng 8(2):191–200. https://doi.org/10.1016/s1672-6529(11)60020-6

Low S, Lim SC, Zhang X, Zhou S, Yeoh LY, Liu YL et al (2017) Development and validation of a predictive model for chronic kidney disease progression in type 2 diabetes mellitus based on a 13-year study in Singapore. Diabetes Res Clin Pract 123:49–54. https://doi.org/10.1016/j.diabres.2016.11.008

Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ et al (2008) Translating the A1C assay into estimated average glucose values. Diabetes Care 31(8):1473–1478. https://doi.org/10.2337/dc08-0545

Patil BM, Joshi RC, Toshniwal D (2010) Hybrid prediction model for type-2 diabetic patients. Expert Syst Appl 37(12):8102–8108. https://doi.org/10.1016/j.eswa.2010.05.078

Purwar A, Singh SK (2015) Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 42(13):5621–5631. https://doi.org/10.1016/j.eswa.2015.02.050

Reed P, Wu Y (2013) Logistic regression for risk factor modelling in stuttering research. J Fluen Disord 38(2):88–101. https://doi.org/10.1016/j.jfludis.2012.09.003

Sahu B, Mishra D (2012) A novel feature selection algorithm using particle swarm optimization for cancer microarray data. Procedia Eng 38:27–31. https://doi.org/10.1016/j.proeng.2012.06.005

Sheik Abdullah A (2012) A data mining model to predict and analyze the events related to coronary heart disease using decision trees with particle swarm optimization for feature selection. Int J Comput Appl 55(8):49–55. https://doi.org/10.5120/8779-2736

Sheik Abdullah A, Selvakumar S, Karthikeyan P, Venkatesh M (2017) Comparing the efficacy of decision tree and its variants using medical data. Indian J Sci Technol 10(18):1–8. https://doi.org/10.17485/ijst/2017/v10i18/111768

Shi Y, Eberhart RC (1998a) A modified particle swarm optimizer. In: IEEE international conference on evolutionary computation proceedings. IEEE World Congress on computational intelligence. http://dx.doi.org/10.1109/icec.1998.699146

Shi Y, Eberhart RC (1998b) Parameter selection in particle swarm optimization. In: Evolutionary programming VII. Springer, Berlin, pp 591–600. http://dx.doi.org/10.1007/bfb0040810

Shi Y, Eberhart RC (1999) Empirical study of particle swarm optimization. In: Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat No 99TH8406), IEEE. http://dx.doi.org/10.1109/cec.1999.785511

Shortliffe EH, Cimino JJ (2014) Biomedical informatics. Springer, London. https://doi.org/10.1007/978-1-4471-4474-8

Smith R, Slenning B (2000) Decision analysis: dealing with uncertainty in diagnostic testing. Prev Vet Med 45(1–2):139–162. https://doi.org/10.1016/s0167-5877(00)00121-5

Steyerberg EW (2009) Clinical prediction models. Statistics for biology and health. Springer, New York. https://doi.org/10.1007/978-0-387-77244-8

Talbi EG (2009) Metaheuristics. Wiley, London. https://doi.org/10.1002/9780470496916

Tang K, Li Z, Luo L, Liu B (2015) Multi-strategy adaptive particle swarm optimization for numerical optimization. Eng Appl Artif Intell 37:9–19. https://doi.org/10.1016/j.engappai.2014.08.002

Van der Zwaan GL, van Dijk SEM, Adriaanse MC, van Marwijk HWJ, van Tulder MW, Pols AD et al (2016) Diagnostic accuracy of the Patient Health Questionnaire-9 for assessment of depression in type II diabetes mellitus and/or coronary heart disease in primary care. J Affect Disord 190:68–74. https://doi.org/10.1016/j.jad.2015.09.045

Wang D, Tan D, Liu L (2017) Particle swarm optimization algorithm: an overview. Soft Comput 22(2):387–408. https://doi.org/10.1007/s00500-016-2474-6

West C, Ploth D, Fonner V, Mbwambo J, Fredrick F, Sweat M (2016) Developing a screening algorithm for type II diabetes mellitus in the resource-limited setting of rural tanzania. Am J Med Sci 351(4):408–415. https://doi.org/10.1016/j.amjms.2016.01.012

Xue B, Zhang M, Browne WN (2014) Particle swarm optimisation for feature selection in classification: novel initialisation and updating mechanisms. Appl Soft Comput 18:261–276. https://doi.org/10.1016/j.asoc.2013.09.018

Yang H, Garibaldi JM (2015) A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform 58:S171–S182. https://doi.org/10.1016/j.jbi.2015.09.006

Zheng T, Xie W, Xu L, He X, Zhang Y, You M et al (2017) A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 97:120–127. https://doi.org/10.1016/j.ijmedinf.2016.09.014

Zyout I, Czajkowska J, Grzegorzek M (2015) Multi-scale textural feature extraction and particle swarm optimization based model selection for false positive reduction in mammography. Comput Med Imaging Graph 46:95–107. https://doi.org/10.1016/j.compmedimag.2015.02.005