Received: October 15, 2020. Revised: January 11, 2021.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Using Hybrid Model of Particle Swarm Optimization and Multi-Layer Perceptron Neural Networks for Classification of Diabetes

Haneen Qteat[1] Mohammed Awad[2]*

[1]Department of Computer Science, Arab American University, Palestine
[2]Department of Computer Systems Engineering, Arab American University, Palestine
* Corresponding author's Email: mohammed.awad@aaup.edu

Abstract: Diabetes mellitus is one of the deadliest and chronic diseases that affect persons who have an increase in their blood glucose levels. Type 1 Diabetes Mellitus "T1DM" is considered one of the most dangerous types of diabetes as it is the reason why diabetes is called the silent killer. Due to the common symptoms of type 1 and 2 diabetes, diabetes specialists face doubts about their diagnosis of the type of diabetes in the patient where the uncertainty about the diagnosis of the disease may lead to delays in controlling the potential complications, especially if they have T1DM. The prime motto of this work is to classify the diabetes types accurately. In this work, we have collected a local Palestinian dataset "DataPal" With the assistance of the Palestinian Diabetes Institute. The "DataPal" dataset was applied using machine learning algorithms to predict diabetes types. The "DataPal" consists of 9 predictors used to diagnose diabetes types. The dataset consists of 314 instances of diabetic females. Thus, our samples were for females with both types 1 and 2 diabetes, aged between 5 and 89 years. The local dataset "DataPal" was preprocessed using the K Nearest Neighbor (KNN) algorithm to fill their missing values, wherein medical diagnosis there is no room for error. The Support Vector Machine (SVM) algorithm was applied to the dataset to select the most optimal features to predict diabetes types. Both the two-fold and four-fold cross-validation methods were applied to the datasets to evaluate the applied models fairly. A hybrid Model Particle Swarm Optimization with Multi-Layer Perceptron Neural Networks (PSO-MLPNNs) uses the PSO evolutionary algorithm to train an MLPNNs and find the optimal weight values of the trained network. The PSO-MLPNNs model was applied to the preprocessed dataset. Then the performance of the model was evaluated using different metrics such as the overall accuracy, recall, specificity, and others. The obtained results show that the proposed PSO-MLPNNs model outperformed all models applied in this work in the classification of diabetes types with an overall accuracy (ACC= 98.73%).

Keywords: Classification, Diabetes mellitus, Diabetes mellitus types, Particle swarm optimization, Multilayer perceptron Neural networks, T1DM, T2DM, Localized diabetes dataset.

1. Introduction

Diabetes mellitus is a chronic disease and one of the 10 deadliest diseases in countries with a strong economy [1]. Diabetes causes many complications that greatly affect human life [2]. Diabetes is one of the main causes of myocardial infarction, diabetic retinopathy, diabetic bone necrosis, kidney failure, amputation, diabetic neuropathy, diabetic coma, and other complications that have not been discovered yet. The number of diabetic people with both types of diabetes has been increasing over the last forty years. According to the World Health Organization [1], diabetes is an epidemic disease that poses a threat to human life. 80% of diabetes deaths are in low-income countries occur in low and middle-income countries [1]. Diabetes caused 4.2 million deaths [3]. Diabetes can be controlled whenever it is detected in the early stages. Also, the correct diagnosis of the type of disease plays a major role in reducing the symptoms of the disease. T1DM is the most dangerous level of diabetes as it is the reason why diabetes is called the silent killer. Diabetes affects people with high blood sugar in the body. There are two different types of diabetes: T1DM and T2DM, which affect people who are unable to produce insulin or are unable to use it to

DOI: 10.22266/ijies2021.0630.02

convert the glucose in the blood into energy [2]. T1DM usually affects young people, mostly under the age of 30. The most common symptom of T1DM is high blood sugar and constant thirst [4]. This type of diabetes cannot be prevented and can only be treated with insulin injections. T2DM is the most common diabetes type and often affects adults. Usually associated with high blood pressure, obesity, atherosclerosis, and other diseases [5].

Usually, doctors and diabetes specialists use some of the tests and factors that help them in the diagnosis of diabetes and its type. Diabetes is usually diagnosed by a person's weight, blood pressure, fasting blood glucose FBS, random blood glucose RBS, the Hemoglobin A1C, and many other tests [4, 6, 7]. To help detect diabetes early, machine learning mechanisms have proven their ability to help doctors make an initial diagnosis and it can also be a source to confirm their diagnoses [8, 9]. To predict diabetes accurately, all attributes that will be used to predict the disease have to contain the correct values without any missing values. The features that are best suited to predict diabetes must be selected. A set of preprocessing steps has to be applied on datasets to fill in the missing values and select the optimal features. To select the optimal features, different methodologies can be applied to any medical dataset to select their optimal attributes such as the genetic algorithm [10], linear discriminant analysis [11], and some of the evolutionary algorithms [12, 13]. It has been proven that Machine learning techniques can make a medical diagnosis, especially diabetes prediction. Support Vector Machine "SVM" with different kernel functions (i.e. Linear, Radial Base [14], Polynomial, and Gaussian) functions, K-Nearest Neighbor "KNN", Discriminant Analysis Classifier "DA", Naive Bayes "NB", Decision Tree "DT" and Random Forest "RF" algorithms are some of the most common machine learning techniques used to predict diabetes.

To classify diabetes types efficiently, in this work we applied various techniques of machine learning including a new enhanced model "PSO- MLPNNs". A local dataset was collected and used to predict diabetes types (i.e. T1DM and T2DM). With the assistance of the Palestinian Diabetes Institute, we have collected a local Palestinian dataset “DataPal”.

In the preprocessing phase, the K-nearest neighbor algorithm [14] was used to estimate the missing values, where, SVM [39] classifier applied for feature selection, K-fold cross-validation has been used to split the datasets into two subsets training and testing. Both two-fold and four-fold cross-validation methods have been applied to validate the performance of the applied models. PSO-MLPNNs proposed a hybrid model that uses one of the most powerful evolutionary algorithms PSO to optimize the weights of MLPNNs. The combination of these two approaches with the tuning of the PSO parameters tends usually to converge the optimization process effectively. The PSO algorithm supposes to be the training method of the MLPNNs model to adjust the weights of the network. To prove the ability of the applied models including the PSO-MLPNNs. A set of performance metrics were used [15, 16]. Confusion matrix, classification accuracy, recalls, specificity is some of the main measurements used to evaluate the performance of the applied models in this work. To prove the efficiency of the PSO-MLPNNs, different machine learning approaches have been applied to classify DataPal as support vector machine SVM, K-nearest neighbor KNN, Decision Tree DT, MLP-BPNNs.

The paper is organized as follows. Section 2 will introduce a set of previous works within the same research field. The "DataPal" dataset description will be shown in section 3. Section 4 will illustrate the algorithmic foundations of the proposed approach. The preprocessing phase, implementation platform, system parameters, and the performance metrics will be presented in Section 5. The experimental Result will be presented and discussed in section 6. Conclusions and future works will be discussed in section 7.

2. Related works

The purpose of this research is the creation of a model to classify and to provide predictive analysis on the diagnosis of diabetes, which allows organizations that provide services to cover health access information on the diagnosis of diabetes mellitus quickly, through using machine learning techniques. The main objective of applying machine learning methods to classify medical datasets is to help the medical specialist and physicians process the non-linear data automatically and find the correct diagnosis [17]. On the other hand, there is no such Palestinian local dataset that has been applied to any machine learning method to predict and classify the types of diabetes. So, there is a need to optimize the diagnosis of diabetes mellitus through an evaluation process of symptomatic features and risk factors using machine learning techniques.

In previous years, many researchers have used different machine learning techniques to predict and diagnose diabetes. Zou Q et.al. [18] used a 12 attributes dataset in diabetes predictions. Random Forest "RF" methodology reached the highest accuracy by up to 80.84% and 77.2% for Luzhou and

PIDD datasets respectively. J. Beschi Raja and others [36] have applied a new proposed methodology to predict type 2 diabetes using both PSO and fuzzy means clustering methodologies. They obtained an overall accuracy up to 95.42% when they have applied their proposed methodology to the PIDD dataset. N. Mohana Sundaram [19] proposed an approach of applying the Elman neural network with MLP neural network.-GAs, PSO, and ACO proved their ability to support the NNs in diabetes predictions as well. Karamath Ateeq and Dr. Gopinath Ganapathy [42] have proposed a modified PSO "MPSO" model to predict diabetes. Both an MLPNNs and a Radial Basis Function Network "RBFN" models were applied to validate their proposed model on both PIDD and US datasets. Where the proposed hybrid MPSO-NNs model was the winner, were achieved 84.2% and 81.8% overall accuracy by applying the US and the PIDD datasets, respectively.

A lot of evolutionary algorithms are used in medical diagnoses in general and in diabetes pattern recognition specifically. Vaishali and others [12] have proposed a hybrid approach consisting of a Fuzzy classifier with Multi-Objective Evolutionary to predict diabetes. The GAs are used as a feature selection classifier to select the most relevant attributes from the PIDD dataset. The model based on rules which are suitable to apply with the critical decision support systems. Using the proposed hybrid model they got an overall accuracy of up to 83%. GAs proposed by D. K. Choubey [10] as a feature selection method as well. The Naive Bayes classification algorithm has been applied to the PIDD dataset to diagnose diabetes. The results show that the Naive Bayes algorithm got 78% overall accuracy. Thus, Ebru Pekel Özmen and Tuncay Özcan [21] have applied the GAs to improve different machine learning methodologies. They have proposed new hybrid methodologies consist of the ANNs and CART classification algorithms and GAs. The hybrid CART-GA and ANN-GA models have been applied to the PIDD dataset, where the highest accuracy was obtained by the proposed CART-GA model by up to 96.05%. SVM, Naïve Bayes, and Decision trees automatic classification algorithms were applied by Deepti Sisodia and, Dilip Singh Sisodia [22] to predict diabetes. The algorithms have been applied and tested to the PIDD dataset. The overall accuracy values obtained by applying the SVM, Naïve Bayes, and Decision trees models are 65.1%, 73.6%, and 76.3%, respectively.

G. Krishnaveni and T. Sudha [14] applied a set of data mining techniques to predict diabetes including, KNN, Discriminant Analysis DA, Naive Bayes NB, and SVM with different kernels functions. They obtained an overall accuracy of up to 76.3% using the Discriminant Analysis DA algorithm. Ashok Kumar Dwivedi [23] has evaluated various machine learning algorithms for predicting diabetes. Classification Tree, ANN, SVM, Naive Bayes "NB", K-NN", and logistic regression are the techniques applied to PIDD to predict diabetes. Both the ANN and SVM models got the highest accuracy by up to 77% and 78% respectively. inyechil Alehegn and others [24] Have applied Naive Net, Decision Stump "DS", and SVM algorithms to validate their proposed model to the PIDD dataset. Their proposed model achieved an overall accuracy of up to 90%.

The PSO evolutionary algorithm has been applied to predict medical diagnosis in general and diabetes in particular. Yuan et al [25] have applied each of the GAs and the PSO to optimize the parameters of the SVM classifier. G. Kranthi Kumar and K. Swathi [26] also proposed the PSO algorithm to adjust the SVM classifier parameters. Then the optimized SVM has been used to classify the PIDD dataset. Alaa Badr Eysa et al [27] have applied two models to predict diabetes. The MLP-BPNN and PSO-NN hybrid approach have been applied to the PIDD dataset to classify diabetes. Where the implementation of both models achieved an accuracy of 77.8% and 88.2% respectively. Sejdinović, and others [28] have proposed an approach to classify the T2DM and Prediabetic instances according to two predictors. A dataset consists of 190 samples was applied by ANNs to classify T2DM and Pre-diabetic using both the FPG and the HbA1c predictors (i.e. Fasting blood sugar and cumulative diabetes during the last three months). They rated 94.1% and 93.3% of cases of Pre-diabetic and T2DM correctly.

In this research, a proposed PSO-MLPNNs model and a set of machine learning techniques will be used to classify diabetes mellitus and its types (i.e.T1DM and T2DM). A Palestinian dataset DataPal has been collected to be used in this work to evaluate applied models in classifying the Diabetes types. In the proposed hybrid model PSO-MLPNNs, PSO which is a global optimization approach will be used to tuning the MLPNN weights. This hybrid model proves its efficacy in the classification of diabetes.

3. Dataset

The DataPal dataset is a Palestinian dataset locally collected from the Palestinian Institute of Diabetes. It consists of 9 predictors used to diagnose diabetes types. The dataset consists of 314 diabetic females both type 1 and types 2 diabetes, aged between 5 and 89 years. Diabetes has many types, where two of the most common types are T1DM and T2DM. Type I diabetes T1DM is the most dangerous level of diabetes mellitus that may cause death. T1DM infects the persons, who do not produce insulin at all, the hormone necessary for the absorption and utilization of glucose. Where this type of diabetes constitutes less than 20% of diabetic people [1]. T1DM is mostly common among young people. Type II diabetes T2DM is less dangerous compared to T1DM. This type affects individuals directly and it common for older people. It occurs when the patients are unable to metabolize glucose [29]. The following are the attributes used to diagnose diabetes types:

- *Age*: The age of the patient is one of the most powerful predictors in predicting diabetes types. Where the T1DM is mostly common among young people while T2DM is common for older people
- *Diabetes Mellitus DM-family history*: Having first and second-degree relatives with diabetes have the effect of transmitting the disease from generation to generation.
- *Pregnancies:* Number of pregnancies. The number of pregnancies affects the determination of the type of sugar because pregnant women are more likely than others to develop diabetes.
- *Body mass index "BMI":* The body fat index according to the height and the weight of patients. BMI has a role in the classification of the type of diabetes. Eq. (1) used to calculate the BMI value [30]:

$$BMI \ = weight \ \ (kg) \div \ height\ (m\ )^2 \quad (1)$$

- *Hypertension "HTN":* Hypertension in family history. Blood pressure has a direct relationship with diabetes, where blood pressure is one of the most important factors affecting diabetes.
- *Ischemic Heart Disease "IHD":* People with both diabetes and cardiovascular disease are more likely to die [5] [31] where both diseases are closely linked to each other. Thus, the genetic factor has a role in the transmission of the disease between generations.
- *Fasting Blood Sugar "FBS":* Blood tests are done after fasting a full night to find out the blood sugar level [32].
- *Human Glycated Hemoglobin A1c "HbA1c":* Test of the average blood sugar associated with hemoglobin during the last two months or three months without the need to fast before the test [32].

9- Blood Pressure "BP": The patient diastolic blood pressure value.

Table 1. The max and min values for the DataPal dataset attributes are shown in Table 1

Attribute	Min Val	Max Val
Age (years)	5	89
DM (The number of diabetic people in the family [0:2])	0	2
Pregnancies (Number of times pregnant)	0	17
BMI (weight in kg/(height in m)^2)	19	40
HTN (The number of people with high blood pressure in the family [0:2])	0	2
IHD (Number of people with cardiovascular disease in the family [0:2])	0	2
FBS (mg/dL)	58	612
BP (mm Hg)	45	110
HbA1c (mmol/L)	5	15

4. Algorithmic foundations

4.1 Artificial neural networks

Artificial neural networks are one of the most widely used classification methods. After collecting the data sets to be classified, they are divided into training and testing datasets so that each of them is applied individually. The optimal weight vectors are obtained by applying the largest set of the dataset which is usually 2/3 of the whole dataset, where the optimal weights will be used to test the trained network with the remaining dataset After completing the network training, it is necessary to calculate the classification error to adjust the network parameters to obtain the lowest error and the highest accuracy.

4.2 Particle swarm optimization "PSO"

Many evolutionary algorithms have been applied to optimize the ANNs to find the optimal weights of the network [34]. PSO is used on the optimization problems of machine learning to train the ANNs and classify variant datasets. In this research, we propose to use a hybrid approach of the MLPNNs model with the PSO to improve the diagnosis of Diabetes Mellitus and its types. The PSO algorithm supposes to be the training method of the MLPNNs model to adjust the weights of the network. PSO algorithm consists of $n = [1,2, \dots n]$ particles fly over a S-dimensional search space, $S = [1,2, \dots S]$ according to the number of input attributes of a dataset. Each particle $x_i$has an initial position x_{il} within the search space S, an initial velocity v_i and a personal best position $P_{best,i}$ determined based on the best value of the fitness function obtained by a particle within a search space. The global best position G_{best} is the position of the particle that has obtained the best fitness value among all particles. Solving an optimization problem means that a set of particles fly over a search space to evaluate the personal and global possible solutions. Where the velocity and the position of each particle are updated according to the best global and personal fitness values. According to the model used in our work, the personal best position $P_{best,i}$ is calculated according to Eq. (2):

$$P_{best,i} = \{P_{best,i}\ , if \quad f(x_i) > P_{best,i}\} \\ else\, \{x_i\ , if \quad f(x_i) \leq P_{best,i}\ \} \qquad (2)$$

Where the new best position could be the current position if the particle has not obtained better fitness value or it is a new position if it is gained a new best fitness value. Then to calculate the global best position G_{best} Eq. (3):

$$G_{best,i} = \{\min(P_{best,i})\, , where\ i \in n = [1{,}2, \ldots n]\,\} \qquad (3)$$

Where it is the position of the particle with minimum error. The following Eq. (4) used to calculate the velocity of the particles:

$$v_{i,new} = w\, v_i + c_1 r_1 (P_{best,i} - x_i) + c_2 r_2 (G_{best,i} - x_i) \qquad (4)$$

Where w is the inertia weight and $(w\, v_i)$ is the inertia component that helps the particles to move in the correct direction. The higher the value of the inertia component the higher the chance to explore the whole search space. Therefore, the inertia component value gives the poise between the effect of the individual component in exploiting the search space and the effect of the social component in exploring the whole search space. The value of the inertia component lies between [0.8 , 1.2] [35], in our model we have supposed that it can be calculated as: $w = rand * 0.35$, where r_1 and r_2 are random values between $[0{,}1]$, $c1$ and $c2$ are coefficients of the individual learning component (i.e. $c_1 r_1 (P_{best,i} - x_i)$) and social learning component (i.e. $c_2 r_2 (G_{best,i} - x_i)$) where c_1 and c_2 usually lies between 0 and 2 [35]. After calculating the velocity of particles the new position they will be determined by Eq. (5):

$$x_{i,new} = x_i + v_{i,new} \qquad (5)$$

Where the new position of the particles in the current position and the velocity of the particle. As for how this evolutionary algorithm-PSO trains neural networks, each particle represents a set of weights (i.e. the neural feed-forward network weights) values which are the dimension of the particle. The n particles fly over the weights search space to minimize the classification error (i.e. maximum fitness) to update the local and global best positions. The training process of the PSO-MLPNNs is shown in Fig. 1.

Step 1: A set of particles with random positions (i.e. initial random weights) is initialized.

Step 2: A MLPNNs is trained using the initialized positions of the PSO algorithm.

Step 3: the learning error is calculated to evaluate the performance of the training process.

Step 4: the velocity and positions of the particles are updated according to the new best local and global positions. Hence the new positions of the particles are the new weights that will be used to train the PSO-MLPNNs again, and so on until the minimum classification error is reached.

Step 5: The new position (i.e. new weights) of each particle is updated by adding the new velocity to the old position value. Where the best global position with minimum classification error is the solution to our optimization problem.

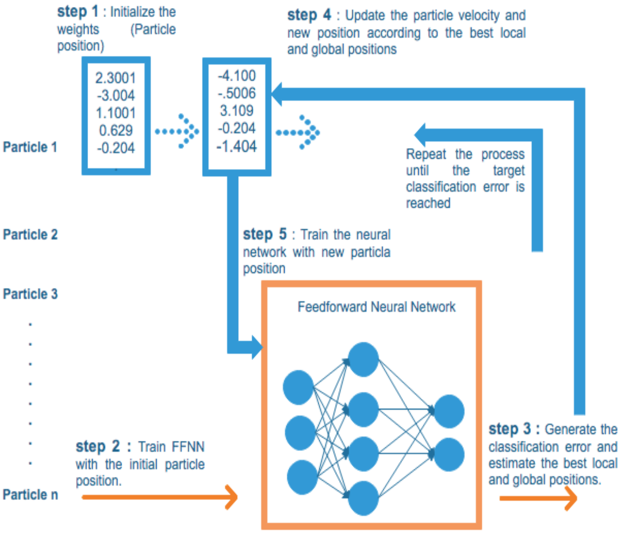
step 1 : Initialize the weights (Particle position)
2.3001
-3.004
1.1001
0.629
-0.204
step 4 : Update the particle velocity and new position according to the best local and global positions
-4.100
-.5006
3.109
-0.204
-1.404
Particle 1
Repeat the process until the target classification error is reached
Particle 2
step 5 : Train the neural network with new particla position
Particle 3
Feedforward Neural Network
step 2 : Train FFNN with the initial particle position.
Particle n
step 3 : Generate the classification error and estimate the best local and global positions.

Figure. 1 The process of training PSO- MLPNNs

The output of each node of the trained MLPNNs is calculated according to Eq. (6):

$$y = \sum_{i=0}^{j}(w_i\, x_i + b_i) \qquad (6)$$

where i € [0,1,...j] and j is # of inputs.

The ANNs input layer has no function to apply to the network inputs, where the inputs forward directly to the hidden layer. But the hidden and the output layers have to apply one or more functions to pass their neurons outputs. The sigmoidal and binary threshold or step activation functions are two of the most commonly applied activation functions [37]. The sigmoid function has been used to activate the output of the MLPNNs hidden layer to estimate the probability of disease occurs as shown in Eq. (8). While the binary threshold or step function has been used to predict the final output of the MLPNNs as shown in Eq. (7).

$$f(y) = (1, if\ y \geq 0)\ else\ (0, if\ y < 0) \qquad (7)$$

$$Y = 1/(1 + exp^{-x}) \qquad (8)$$

5. Implementation

The Applied models in this work are simulated using the Matlab R2019a platform. We have applied a set of machine learning algorithms that have been applied in various previous studies to validate the proposed PSO-MLPNNs model. SVM with different kernel functions (i.e. Linear, Radial Base, Polynomial, and Gaussian) [14], KNN, Discriminant Analysis Classifier “DA” [11], Naive Bayes “NB” [9], decision tree “DT” and Random Forest “RF” [42], where the algorithms have been used to prove the ability of the PSO-MLPNNs model in classifying diabetes.

5.1 System parameters

To train a feed-forward neural network we have to adjust the number of hidden neurons in hidden layers of the network. In this work, the hidden neurons were adjusted by try and error way to gain the best pattern recognition results. Table 2 summarizes the optimal values of the following PSO parameters used in the proposed PSO-MLPNNs model in this work:

- *The coefficients of the individual learning component $c1$and $c2$:* are the components that allow particles to exploiting and exploring the search space of the PSO to search for a solution. The optimal values for both components were found to be 2 by increasing their values in the try and error approach.
- *The number of Population or Particles:* the number of particles has been adjusted to be 30 particles, the smallest number of particles needed to search the search space for a solution to keep the model within a stable generalization status. The small number of particles makes the model gets rid of its slow state if a large number of molecules are used [35].
- *The inertia weight component:* the component that helps the particles to move in the correct direction. The higher the value of the inertia component the higher the chance to explore the whole search space. Therefore, the inertia component value gives the balance between the effect of the individual component in exploiting the search space and the effect of the social component in exploring the whole search space. The value of the inertia component lies between $[0.8: 1.2]$ [35], in our model we have supposed that it can be calculated: as$= rand \times 0.35$.
- *The initial velocity v:* the particles start searching the search space with an initial velocity v. In our model, we have supposed that the initial velocity v can be calculated: as$= 0.15 \times x_i$, where x_i is the initial position of particles.
- *The search space boundaries (i.e. the Lower boundary and the upper boundary) LB and UB :* search spaces should have boundaries to prevent the particles from boundary violations and to stay within the determined boundaries. In this work, each of the lower and upper boundaries has been adjusted to be -1.5 and 1.5 respectively. They were selected based on the results obtained by applying different values and choose the optimal among them.

Table 2. The PSO-MLPNNs model optimal parameters

Parameter	**The optimal value**
$c1$	2
$c2$	2
Population number	30
w	$rand \times 0.35$
v	$0.15 \times x_i$
LB	-1.5
UB	1.5
Activation function	Sigmoidal function \ hidden layer. The Binary Step \ output layer.

5.2 Feature selection

Some features in the applied dataset may be irrelevant based on specialists in the field of diabetes. So, optimal features must be selected. In this work, we have applied the SVM method to select the relevant features. By applying feature selection to the "DataPal" local data set, some features that contain duplicate or non-useful records for the classification of diabetes and its types were excluded.

5.3 Data partitioning

Datasets divided into groups to evaluate the classification model performance. The K-Fold cross-validation methodology will be used to evaluate the applied models. Where the dataset samples are separated into K-Folds randomly to train and test the groups through K number of iterations. At each iteration, one of the K folds will represent the testing dataset and the rest will be used to train the model. This method is optimal in the case of limited data. The generated folds are evaluated alternately wherein the first iteration the contents of the first fold for testing the data set and the rest for training and vice versa in the second Iteration. We use both two-fold and four-fold cross-validation because the distribution of our limited dataset.

5.4 Performance metrics

The classification accuracy, misclassification rate, sensitivity, specificity, precision, geometric mean, and F-Measure metrics were used to evaluate the three models' performance [15] [16]. These metrics are used with the models that typically apply the imbalanced datasets in their evaluations.
The Accuracy of pattern recognition models is calculated according Eq. (9):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \times 100\ \% \qquad (9)$$

Precision**:** is the percentage of samples that have been predicted as positive class correctly to all samples that have been predicted as a positive class as shown in Eq. (10):

$$Precision = \frac{TP}{TP+FP} \times 100\ \% \qquad (10)$$

Sensitivity or Recall: this is the percentage of samples that have been predicted as positive class correctly to all samples that have been predicted as positive class correctly and incorrectly as a negative class. Is the measurement used to evaluate the accuracy of positive cases, as shown in Eq. (11)

$$Sensitivity = \frac{TP}{TP+FN} \times 100\ \% \qquad (11)$$

Specificity: is the percentage of samples that have been predicted as negative class correctly to all samples that have been predicted as negative class correctly and incorrectly as a positive class. Is the measurement used to evaluate the accuracy of negative cases, as shown in Eq. (12):

$$Specificity = \frac{TN}{TN+FP} \times 100\ \% \qquad (12)$$

Where True Positive TP, False Positive FP, True Negative TN, and False Negative FN.

6. Results and discussions

In this section, we compare and discuss the results of our experiments using the applied models. The DataPal dataset is divided using two-fold and four-fold cross-validation into two sub-datasets to train and test the applied models. The SVM with different kernels [14, 39, and 41] which is considered as one of

Table 3. The results of applying different machine learning models using two-fold cross-validation

Approach	**Performance Metrics**		
	Accuracy (%)	Sensitivity (%)	Specificity (%)
MLP-NN [38]	97.15%	99.23%	87.03%
Linear-SVM [14]	96.17%	97.32%	90.38%
Polynomial-SVM [39]	94.9%	96.92%	85.18%
Gaussian-SVM [41]	95.2%	96.97%	88.23%
RBF-SVM [14]	95.2%	96.97%	88.23%
K-NN [41]	93.63%	95.11%	85.41%
DA [41]	96.49%	97.7%	90.56%
NB [40]	96.81%	98.46%	92.3%
DT [41]	95.22%	96.22%	89.79%
RF [42]	95.22%	96.57%	88.23%
PSO-MLPNNs	**98.73%**	**99.24%**	**94.45%**

Received: October 15, 2020. Revised: January 11, 2021.

Table 4. The results of applying different machine learning models using four-fold cross-validation

Approach	Performance Metrics		
	Accuracy (%)	Sensitivity (%)	Specificity (%)
MLP-NN [38]	96.7%	98.97%	85.85%
Linear-SVM [14]	96.18%	97.69%	88.88%
Polynomial-SVM [39]	95.23%	97.66%	84.21%
Gaussian-SVM [41]	95.87%	97.31%	88.67%
RBF-SVM [14]	95.87%	97.31%	88.67%
K-NN [41]	93.62%	96.96%	79.06%
DA [41]	96.81%	97.70%	92.30%
NB [40]	96.81%	97.70%	92.30%
DT [41]	93.94%	97.25%	79.66%
RF [42]	96.17%	97.32%	90.38%
PSO-MLPNNs	97.77%	99.48%	93.12%

the most powerful machine learning algorithms has been applied to the DataPal dataset.

It can be noticed from Table 3 that the Linear-SVM model was accurate in classifying the types of diabetes by up to 96.17%. But it was ineffective in predicting T1DM, which is the most dangerous type of diabetes. Both KNN and DA algorithms were able to predict the types of diabetes by up to 93.63% and 96.49% respectively. Therefore, both KNN and DA were unable to predict T1DM efficiently. Each of the NB, RF, and DT has obtained convergent results in predicting the “DataPal” dataset by up to 96.81%, 95.22%, and 95.22 respectively.

The proposed PSO- MLPNNs model was the most accurate model with classification accuracy up to 98.73%. The PSO- MLPNNs model was the best in T2DM and T1DM predictions by up to 99.48% and 94.45% respectively as shown in Table 3 and Table 4 that summarize the results obtained by applying different machine learning methodologies using two-fold and four-fold cross-validation methodologies.

As shown in Tables 3 and 4, the hybrid model of PSO-MLPNNs outperforms the other known machine learning classification algorithms in classification accuracy, sensitivity, and specificity. The results show that based on various features used in the literature, such as the number of pregnancy rates, diastolic blood pressure, body mass index, age, etc.

The percentage of successes of the hybrid model is very high, it’s clear that the proposed model found

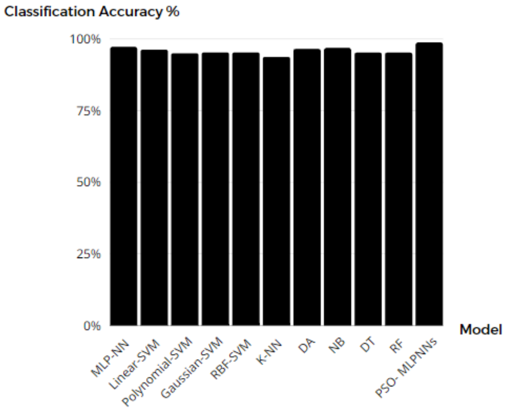

Figure. 2 Accuracy values comparison using two-fold cross-validation

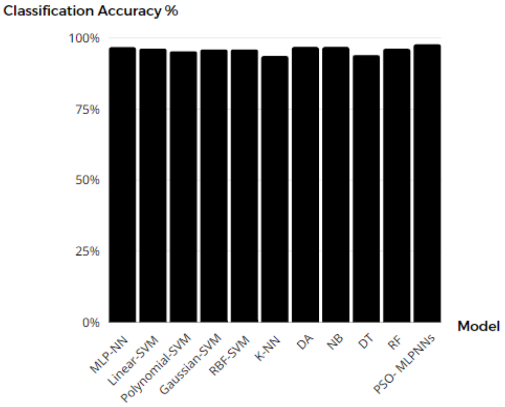

Figure. 3 Accuracy values comparison using four-fold cross-validation

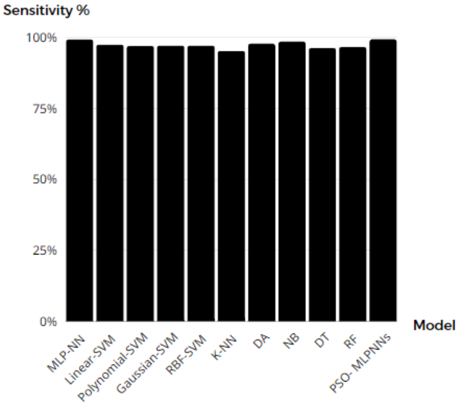

Figure. 4 Sensitivity values comparison using two-fold cross-validation

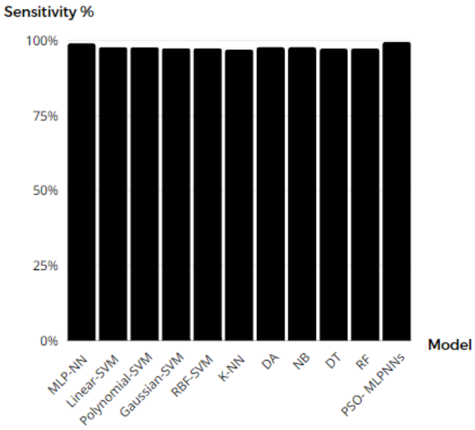

Figure. 5 The sensitivity values comparison using two-fold cross-validation

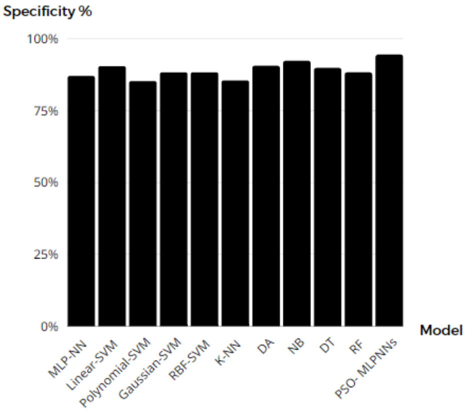

Figure. 6 Specificity values comparison using two-fold cross-validation

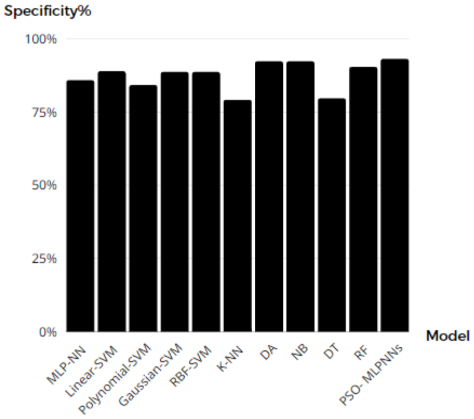

Figure. 7 Specificity values comparison using two-fold cross-validation

better results to diagnose diabetes. This finding further facilitates the detection of the disease by contributing to its prevention Fig. 1 and 3 illustrate the classification accuracy of applying the PSO-MLPNNs model with two-fold and four-fold cross-validation methodologies.

Fig. 2 and 4 show the ability of our model in classifying T2DM by up to 99.48%.

Fig. 3 and 6 show how our model outperforms other models in predicting the most serious types of diabetes T1DM.

7. Conclusions

Taking into account all risks of diabetes, early stages and types of disease must be detected early. Diabetes can be controlled whenever it is detected in the early stages. Also, the correct diagnosis of the type of disease plays a major role in reducing the symptoms of the disease. T1DM is the most dangerous level of diabetes as it is the reason why diabetes is called the silent killer. This type of diabetes cannot be prevented and can only be treated with insulin injections. So, the diagnosis of the type of diabetes must be very accurate.

This research aimed to determine effective variables and their impact on diabetes and estimating a neural network hybrid model with PSO to Predict and classify diabetes types. Thus, in this work, we have collected a local Palestinian dataset "DataPal" With the assistance of the Palestinian Diabetes Institute. For the first time, a Palestinian dataset was applied using machine learning algorithms to predict diabetes types. The collected dataset was used to train the proposed PSO-MLPNNs model in addition to a set of machine learning algorithms. The parameters of the PSO optimization algorithms were optimized. Then the optimized PSO algorithm has been used to adjust the MLPNNs weights values. In conclusion, The PSO-MLPNNs model has proved its ability in predicting each of T1DM and T2DM with an accuracy of 98.73%. The PSO-MLPNNs model has outperformed each of the MLPNNs, SVM, K-NN, DT, DA, NB, and RF algorithms in classifying T1DM and T2DM. Each of the MLP-BPNN and the SVM have gained an accuracy of 97.15% and 96.18% respectively, the highest accuracy percentages compared to the other models applied in this work (i.e. K-NN, DT, DA, NB, and RF). Given the results obtained, the PSO-MLPNNs classifier to identify diabetic people is proposed as a useful tool to help the diabetes specialist in the early detection of the disease and to confirm their diagnosis. In future, we plan to develop a medical application that will help people with a family history of diabetes who suspects they are based on linking machine learning algorithms to provide them with preventive treatments and advice. The Fuzzy-Rule Based system will be used with the PSO-MLPNNs model to predict diabetes and give preventive treatments automatically. Also, developing a prediction and classification system with real-time remote monitoring applications, measure the effective cost of implementing them, and the impact they can have on people's lives.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Author 1: Collect the data, make contributions to conception and design, analysis, and interpretation of data.

Author 2: review and analyze the manuscript, participate in drafting the article, or revising it critically.

References

[1] Who.int. (2019). Diabetes. [online] Available at: http://www.who.int/en/news-room/fact-sheets/detail/diabetes [Accessed 31 Oct.2019].

[2] D. Mauricio, N. Alonso, and M. Gratacòs, “Chronic Diabetes Complications: The Need to Move beyond Classical Concepts”, *Trends in Endocrinology & Metabolism Journal*, Vo1. 31, No. 4, pp. 287-295, 2020.

[3] A. Wako, S. Belay, Y. Feleke, and T. Kebede, “Assessment of risk for severe hypoglycemia among adults with IDDM: validation of the low blood glucose index”, *Journal of Diabetes and Metabolism*, Vol. 21, No. 11, 2017.

[4] Bruzda-Zwiech, J. Szczepańska, and R. Zwiech, “Xerostomia, thirst, sodium gradient and inter-dialytic weight gain in hemodialysis diabetic vs. non-diabetic patients”, *Medicina Oral Patología Oral y Cirugia Bucal*, Vol. 23, No. 4, pp. 406-412, 2018.

[5] M. A. Said, N. Verweij, and P. van der Harst, “Associations of Combined Genetic and Lifestyle Risks with Incident Cardiovascular Disease and Diabetes in the UK Biobank Study”, *JAMA Cardiology,* Vol. 3, No. 8, pp. 693-702, 2018.

[6] E. Barry, S. Roberts, J. Oke, S. Vijayaraghavan., R. Normansell, and T. Greenhalgh, “Efficacy and effectiveness of screen and treat policies in prevention of type 2 diabetes: systematic review and meta-analysis of screening tests and interventions”, *BMJ*, 2017.

[7] M. J. L. Verhulst, B. G. Loos, V. E. A. Gerdes, and W. J. Teeuw, “Evaluating All Potential Oral Complications of Diabetes Mellitus”, *Front Endocrinol (Lausanne)*, Vol. 10, No. 56, p. 56, 2019.

[8] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 104-116, 2017.

[9] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project”, Plos One, Vol. 12, No. 7, 2017.

[10] D. K. Choubey, S. Paul, and S. Kumar, “Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection”, In: *Proc. of International Conf. on Communication and Computing System*, London, pp. 451–455, 2017.

[11] Kolukisa, H. Hacilar, G. Goy, M. Kus, B. Bakir-Gungor, A. Aral, and V. C. Gungor, “Evaluation of Classification Algorithms, Linear Discriminant Analysis and a New Hybrid Feature Selection Methodology for the Diagnosis of Coronary Artery Disease”, In*: Proc. of International Conf. on Big Data (Big Data),* Seattle, Westin Seattle, USA, pp. 2232-2238, 2018.

[12] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, “Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset”, In: *Proc. of International Conf. on Computing Networking and Informatics*, Nigiria, Lagos, pp. 1-5, 2017.

[13] A. Herliana, T. Arifin, S. Susanti, and A.B. Hikmah, “Feature Selection of Diabetic Retinopathy Disease Using Particle Swarm Optimization and Neural Network”, In: *Proc. of International Conf. on Cyber and IT Service Management*, Parapat, Indonesia, pp. 1-4, 2018.

[14] G. Krishnaveni, T. Sudha, “A novel technique to predict diabetic disease using data mining – classification techniques”, *International Journal of Advanced Scientific Technologies, Engineering and Management Sciences*, Vol. 3, No. 1, 2017.

[15] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix”, *Pattern Recognition*, Vol. 91, pp. 216–231, 2019.

[16] J. Akosa, “Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data”, In: *Proc. of the SAS Global Forum,* Orland, Florida, 2017.

[17] Saru, S., and S. Subashree. “Analysis and prediction of diabetes using machine learning”, *International Journal of Emerging Technology and Innovative Engineering* 5.4 *2019.*

[18] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Predicting Diabetes Mellitus with Machine Learning Techniques”, *Journal of Frontiers in Genetics*, Vol. 9, pp. 728-734, 2018.

[19] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 104–116, 2017.

[20] X. Xiong, RX. Zhang, Y. Bi, WH. Zhou, Y. Yu, and DL. Zhu, “Machine Learning Models in Type 2 Diabetes Risk Prediction: Results from a Cross-sectional Retrospective Study in Chinese Adults”, *Current Medical Science*, Vol. 39, No.4, pp. 582–588, 2019.

[21] E. Pekel Özmen, and T. Özcan, “Diagnosis of Diabetes Mellitus using Artificial Neural Network and Classification and Regression Tree optimized with Genetic Algorithm”, *Journal of Forecasting*, Vol. 39, No. 4, pp. 661-670.

[22] Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms”, *Journal of Procedia Computer Science*, Vol. 132, pp. 1578–1585, 2018.

[23] A. K. Dwivedi, “Analysis of computational intelligence techniques for diabetes mellitus prediction”, *Journal of Neural Computing and Applications*, Vol. 30, No. 13, 2018.

[24] M. Alehegn, R. Joshi, and P. Mulay, “Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm”, *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 9, pp. 871-878, 2018.

[25] N. Farooqui, Ritika, and A. Tyagi, “Prediction Model for Diabetes Mellitus Using Machine Learning Techniques”, *International Journal of Computer Sciences and Engineering*, Vol. 6, No. 3, pp. 292–296, 2018.

[26] G. Kranthi Kumar and K. Swathi, “PERFORMANCE IMPROVEMENT APPROACH FOR DIABETES DISEASE PREDICTION”, *International Journal of Computer Application*, Vol. 7, No. 1, pp.2250-1797, 2017.

[27] A. B. Eysa and S. Kurnaz, “Diabetes Diagnosis Using Machine Learning”, *International Journal of Computer Science and Mobile Computing*, Vol. 8, No. 3, pp. 36-41, 2019.

[28] D. Sejdinović, L. Gurbeta, A. Badnjević, M. Malenica, T. Dujić, A. Čaušević, ans L. D. Mehmedović, “CLASSIFICATION OF PREDIABETES AND TYPE 2 DIABETES USING ARTIFICIAL NEURAL NETWORK”, In: *Proc. of International Conf. on Medical and Biological Engineering*, Vol. 62, pp. 685–689, 2017.

[29] L. M. Goff, “Ethnicity and Type 2 diabetes in the UK”, *Diabet Med*, Vol. 36, No. 8, pp.927-938, 2019.

[30] M. Ji, S. Zhang, and R. An, “Effectiveness of A Body Shape Index (ABSI) in predicting chronic diseases and mortality: a systematic review and meta-analysis”, *Obesity Reviews*, Vol. 19, No. 5, pp. 737–759, 2018.

[31] J. Al-Tu’ma, B. A. Joda, and R. A. Al-Yassiry, “Assessment of Vascular Endothelial Growth Factor-A and Insulin Resistance in Sera of Ischemic Heart Diseases with Type-II Diabetic Patients”, *Indian Journal of Natural Sciences*, Vol. 9, No. 51, pp. 976-997, 2018.

[32] Diabetes. (2018, August 8). Retrieved December 15, 2020, from https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-203 71451#targetText=Fasting blood sugar test. &targetText=A fasting blood sugar level, separate tests, you have diabetes.

[33] Diabetes. (2018, August 8). Retrieved December 15, 2020, from https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451.

[34] K. O. Stanley, J. Clune, J. Lehman, and others, “Designing neural networks through neuroevolution”, *Nat Mach Intell* 1, pp. 24–35, 2019.

[35] K. Ateeq and G. Ganapathy, “The novel hybrid Modified Particle Swarm Optimization - Neural Network (MPSONN) Algorithm for classifying the Diabetes”, *International Journal of Computational Intelligence Research*, Vol. 13, No. 4, pp. 595-614, 2017.

[36] J. B. Raja, and S. C. Pandian, “PSO-FCM Based Data Mining Model to Predict Diabetic Disease”, *Computer Methods and Programs in Biomedicine*, 2020.

[37] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, “Activation functions: Comparison of trends in practice and research for deep learning”, arXiv, 2018.

[38] S. K. Mohapatra, J. K. Swain, and M. N. Mohanty, “Detection of Diabetes Using Multilayer Perceptron”, In: *Proc. of*

International Conf. On Intelligent Computing and Applications, pp.109–116, 2019.

[39] P. Verma, I. Kaur, and J. Kaur, “Novel Approach of Diabetes Disease Classification by Support Vector Machine with RBF Kernel”, *International Journal of Advance Research, Ideas and Innovations in Technology*, Vol. 3, No. 1, 2017.

[40] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine learning and data mining methods in diabetes research”, *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 104-116, 2017.

[41] A. Al-Zebari and A. Sengur, “Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection”, In: *Proc. of International Conf. on Informatics and Software Engineering Conf. (UBMYK)*, pp. 1-4, 2019.

[42] S. Benbelkacem and B. Atmani, “Random Forests for Diabetes Diagnosis”, In: P*roc. of International Conf. on Computer and Information Sciences (ICCIS)*, pp. 1-4, 2019.