International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Data Classification of Early-Stage Diabetes Risk Prediction Datasets and Analysis of Algorithm Performance Using Feature Extraction Methods and Machine Learning Techniques

Ali Yasar[1,*]

Submitted: 17/10/2021 **Accepted : 24/12/2021**

Abstract: Diabetes is one of the more common diseases in the world today and is one which also plays a role in the development of many other critical or terminal illnesses such as heart diseases, coronary diseases, eye diseases, kidney diseases, and even nerve damage. Thus, early diagnosis is of great importance. With the development of machine learning techniques and artificial intelligence, the estimation of disease risks has started to be widely accepted and applied by researchers and medical doctors. In this study, a machine learning technique was proposed for the prognosis of early onset diabetes. An interface was designed using the MATLAB graphical user interface (GUI). The wrapper-based Particle Swarm Optimization (PSO), Tree Seed Algorithm (TSA), Crow Search Algorithm (CSA), Slime Mould Algorithm (SMA), and Artificial Bee Colony (ABC) algorithms were used to reduce and select the required input attributes. The results obtained with these algorithms were compared by using conventional machine learning algorithms such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K Nearest Neighbor (kNN) and Feed Forward Neural Networks (FFNN). 16 features used in the diagnosis of diabetes, od the wrapper-based feature selection and feature reduction methods 10 features with PSO method, 9 features with TSA method, 13 features with CSA method, 6 features with SMA method and 8 features with ABC method has been determined. The features determined by each respective method were then classified using machine learning algorithms. All combinations have been tried and these are the results of the best five combinations on the results, methods displayed the best classification performances with success rates of PSO + SVM = 97.5, TSA + SVM = 96.15, CSA + FFNN = 99.04, SMA + FFNN = 94.23, and ABC + SVM = 96.73 respectively.

Keywords: *Classification, CSA, FFNN, Diabetes, Optimization,*

This is an open access article under the CC BY-SA 4.0 license.
(https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Diabetes is a disease that develops when the pancreas does not produce enough of the hormone insulin or when the insulin produced cannot be used effectively and can cause serious complications. The pancreas, an organ located in the upper left abdomen, has 2 fundamental functions: an endocrine function and an exocrine function [1]. Diabetes is caused when the pancreatic beta cells function improperly, which can thus lead to various organ diseases such as impaired renal function, risk of heart disease, hypertension, eye damage that can lead to visual disturbances, nerve damage, and glaucoma [2]. The World Health Organization (WHO) stated in a 2018 report that diabetes is one of the fastest growing chronic life-threatening diseases, affecting approximately 422 million people worldwide. Increasing morbidity in recent years indicates that the number of diabetic patients in the world will reach 642 million by 2040, which means that one in ten adults will have diabetes in the future [3]. There are three types of diabetes: Type 1 diabetes is caused by the failure of the pancreas to create enough insulin due to lack of beta cells; Type 2 diabetes occurs due to insulin resistance, when cells are unable to respond to insulin; the final type is gestational diabetes, which occurs when a pregnant woman with no history of diabetes has high blood glucose volume during pregnancy [4]. Diabetes mellitus, more widely known as simply diabetes, is a disease that occurs with the increase of sugar in the blood over a long period of time, leading to metabolic disorders [5]. Symptoms of diabetes are generally increased hunger, increased thirst, and frequent urination. If neglected, diabetes can lead to more severe long-term health problems such as cardiovascular disease, chronic kidney disorder, stroke, foot ulcers, nerve injury, cognitive impairment and vision loss. [4, 6, 7].

In parallel with advancing technology and developed living conditions, diabetes has also gained prominence, affecting the daily lives of millions. Quick and accurate diagnosis and analysis of early diabetes is thus of great importance and is a matter worth studying. Choosing the correct classification method in machine learning studies is a critical issue in this regard. There are different parameters used in the classification methods for the prognosis of diabetes; to this respect, it was observed that effective results could not be obtained by means of many methods available in the literature [8-14]. Many methods such as Artificial Neural Networks (ANN), DTs, and SVMs have been proposed for the diagnosis of diabetes by using machine learning methods [15]. Islam et al. subjected the dataset used in their study to classification with the Naive Bayes (NB) Algorithm, Logistic Regression (LR) Algorithm and RF Algorithm with 10-fold cross validation and concluded that they obtained the most successful result with the RF method in this dataset [16]. Faniqul et al. conducted their study on the same dataset and applied four algorithms: NB, DT, LR, and

[1] *Department of Mechatronic Engineering, Faculty of Technology, Selcuk University, Konya, Türkiye, ORCID ID: 0000-0001-9012-7950*
** Corresponding Author Email: aliyasar@selcuk.edu.tr*

RF. They calculated the performance of each classifier and found that the most successful method was RF with 10-fold cross validation with 97.4% accuracy [16, 17]. Zou et al. reduced the data size with principal component analysis (PCA) in feature extraction methods using random data from 68,994 patients obtained from a hospital in Luzhou, China. Using the obtained features, they achieved 80.84% accuracy with RF [3]. Yue et al. used the quantum particle swarm optimization (QPSO) algorithm and the weighted least squares support vector machine (WLS-SVM) for the prognosis of Type 2 diabetes [18]. In another study, seven features were used: diabetes risk factor, age, gender, body mass index (BMI), diabetes history in the family, blood pressure, duration of diabetes, and blood glucose level. NB and C4.5 decision tree-based classification methods and k-NN clustering techniques were used for the analysis of this data set [19]. In a study conducted by Ahmed (2016), the J48 algorithm was modeled using the WEKA application in data mining methods on approximately 318 Type 2 diabetic medical data collected from the Jabir Abu Eliz Diabetic Center (JADC) in Sudan and achieved a classification accuracy of 70.8% [20]. In another study conducted by Nurjuhan et al. (2021), a classification using feature selection algorithms in 2 different diabetes datasets (Pima Indian and Sylhet Diabetes Hospital) was made in which the researchers achieved an accuracy ratio of 97.5% and 77.7% for the data obtained from Sylhet Diabetes Hospital with the highest LR regarding the Pima Indian diabetes data. Emon et al. tested the Sylhet Diabetes Hospital dataset with data from 130 patients and their classifiers and stated that they reached the highest accuracy rate of 98% using the RF method [21]. In his study, Taser achieved 98.65% accuracy by applying the Adaboost method on the Naive Bayes Tree (NBTree) of the Sylhet Diabetes Hospital dataset [22].

Patient and non-patient data obtained from 520 samples collected via direct questionnaires from Sylhet Diabetes Hospital in Sylhet, Bangladesh was used in the present study. The aim of this study is to benefit from the given features, to design a prediction algorithm using machine learning, and to find the most appropriate classification algorithm to obtain the closest result compared to clinical results. In this study, machine learning methods were used for the prognosis of early diabetes and the anticipated results were obtained. The decision tree is a popular machine learning method in the medical field as it has strong classification power. Random forests are also widely used as they can produce many decision trees. However, neural networks have recently become more prevalent machine learning methods as they perform better in many ways [3]. Feature selection was performed in this study using the PSO, TSA, CSA, SMA, and ABC algorithms for the prognosis of diabetes; then, the classification performances were analysed using SVM, DT, FFNN, RF, and kNN methods in the classification process.

2. Materials and Methods

2.1. The Dataset

The dataset used in this research was obtained from 520 patients from Sylhet Diabetes Hospital via direct questionnaires conducted under doctor supervision [23]. The dataset consisted of 200 healthy individual and 320 diabetic patients. Physical data are given in the following 16 physical examination data and classification results in Figure 1.

These datasets:

Age: how old the patient is (20-65).

Gender: Shows the patient's gender information (Male/Female).

Polyuria: Production of abnormally large volumes of dilute urine. (Yes/No).

Itching: an uncomfortable sensation on the skin that causes a desire to scratch. (Yes/No)

Irritability: the quality or state of being irritable. (Yes/No)

Delayed healing: It has been defined as 'healing that takes longer than anticipated, given appropriate therapy (Yes/No)

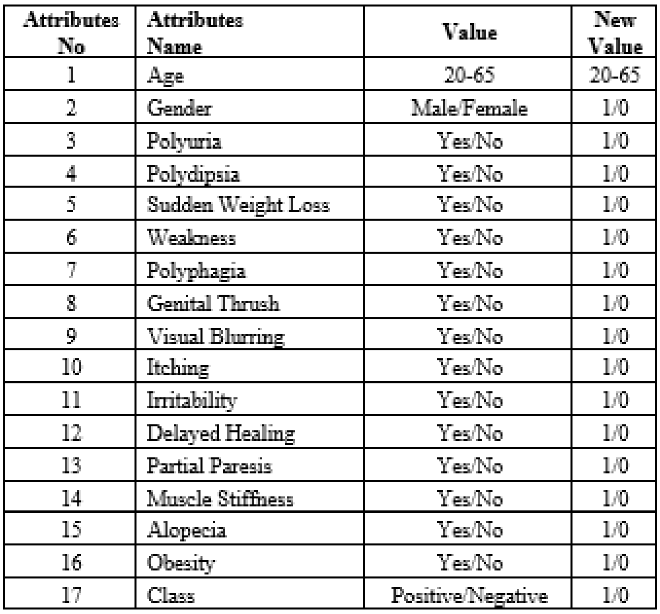
Partial Paresis involves the weakening of a muscle or group of muscles. It may also be referred to as partial or mild paralysis.

Muscle stiffness is when your muscles feel tight and you find it more difficult to move than you usually do, especially after rest. You may also have muscle pains, cramping, and discomfort. (Yes/No)

Alopecia is a condition in which hair is lost from some or all areas of the body. (Yes/No)

Obesity is defined as abnormal or excessive fat accumulation that presents a risk to health. (Yes/No)

Class: Diabetes or Non-Diabetes (Positive / Negative)



Attributes No	Attributes Name	Value	New Value
1	Age	20-65	20-65
2	Gender	Male/Female	1/0
3	Polyuria	Yes/No	1/0
4	Polydipsia	Yes/No	1/0
5	Sudden Weight Loss	Yes/No	1/0
6	Weakness	Yes/No	1/0
7	Polyphagia	Yes/No	1/0
8	Genital Thrush	Yes/No	1/0
9	Visual Blurring	Yes/No	1/0
10	Itching	Yes/No	1/0
11	Irritability	Yes/No	1/0
12	Delayed Healing	Yes/No	1/0
13	Partial Paresis	Yes/No	1/0
14	Muscle Stiffness	Yes/No	1/0
15	Alopecia	Yes/No	1/0
16	Obesity	Yes/No	1/0
17	Class	Positive/Negative	1/0

Figure 1. Features of Dataset

2.2. Data Pre-processing

In order to be used in classification problems, data should be used continuously. Based on this, the present data were transformed in such a way that they could be used for data pre-processing. Figure 1 provides features of the relevant dataset.

2.3. Feature Selection (FS)

"Relevant" identification of a feature is a difficult issue due to the complex (two-way, three-way, or multi-faceted) interactions between features. A feature may become relevant or irrelevant when used in combination with other features; therefore, an optimal property subset should contain complementary properties that provide various properties of classes. The feature selection process locates a small feature subset in a particular large feature data set and has been of interest for many years [24, 25]. Feature selection in classification is to find a subset of the relevant features so that the dimensionality of the data can be reduced and the learning/classification process can be accelerated while the overall classification performance can be maintained or improved [25, 26]. Initiation, search strategy, evaluation criteria, and retention criteria stages are factors that affect the performance of a feature selection algorithm. Various algorithms have been proposed in order to address feature selection problems; however, they still remain a challenge. In this section, five feature selection approaches (PSO, TSA, CSA, SMA, and ABC) are discussed.

2.3.1. PSO Feature Selection

PSO imitates the social behavior of birds that live in flocks [27]. PSO is a meta-heuristic search algorithm that simulates the movements of a flock of birds to find a food source. Each parcel of the swarm represents a candidate solution flying in the multidimensional search space. A parcel uses the best position discovered by itself and its neighbors to move towards an optimum solution [28]. The algorithm was developed using computer simulations and various interpretations. PSO uses various substances (parcels) that forms a flock. This flock navigates the search area to locate the best possible solution. With respect to each parcel in the search area, the flock alters the experience of flying and the "flying" of other parcels to imitate the experience of flying. PSO is initiated by randomly generated parcels and their velocity, which indicates the search speed. Then, the particles are evaluated in terms of relevance, as in the way the GA algorithm works. Two main tests follow such evaluations. The first compares the experience of a parcel called personal best (pbest) to itself. The second test, the global best (gbest), compares the relevance of the parcel to the overall flock experience. Conducting these two tests ensures the protection of the best parcel. These processes are performed until the retention criteria is determined [29]. Figure 2 provides the PSO flow chart [30].

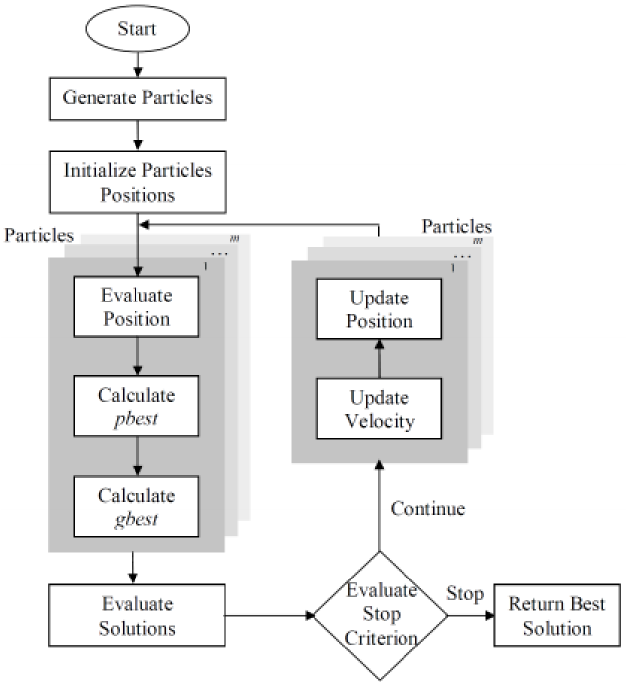


Figure 2. Diagram of PSO

2.3.2. TSA Feature Selection

The population-based Tree-Seed algorithm (TSA) was proposed by Kiran to solve continuous optimization problems [31]. TSA refers to solutions by mimicking the relationship between trees and their seeds in nature [32]. Assuming that the surface of these trees is a search area for the optimization problem, the location of trees and their seeds can be considered possible solutions [33]. First, trees are created using Equation (1).

$$T_{i,j} = Low_j + r_{i,j}(High_j - Low_j) \tag{1}$$

where Ti,j is the jth dimension of the ith tree, ith is a uniformly random number in the range of [0,1], Lowj is the lower bound of the jth dimension, Highj is the higher bound of the jth dimension. Two search equations were designed for this process. The important point here is that the equation is chosen to generate a new seed location controlled by a parameter of the method called the search propensity Search Tendency (ST) in the range of [0, 1]. All iteration's seeds are created by using Equation (2) or (3).

$$S_{i,j} = T_j + \alpha_{i,j}(Best_j - T_{r,j}) \tag{2}$$

$$S_{i,j} = T_{i,j} + \alpha_{i,j}(T_{i,j} - T_{r,j}) \tag{3}$$

Equation (2) determines the location of the tree from which the seed will be produced and the best location of the tree population for the relevant tree. This search equation also improves the local search or concentration capability of the proposed algorithm. The third update rule, or Equation (3), uses two different tree locations to generate a new seed for the tree. The flow chart of the TSA is shown in Figure 3 [34].

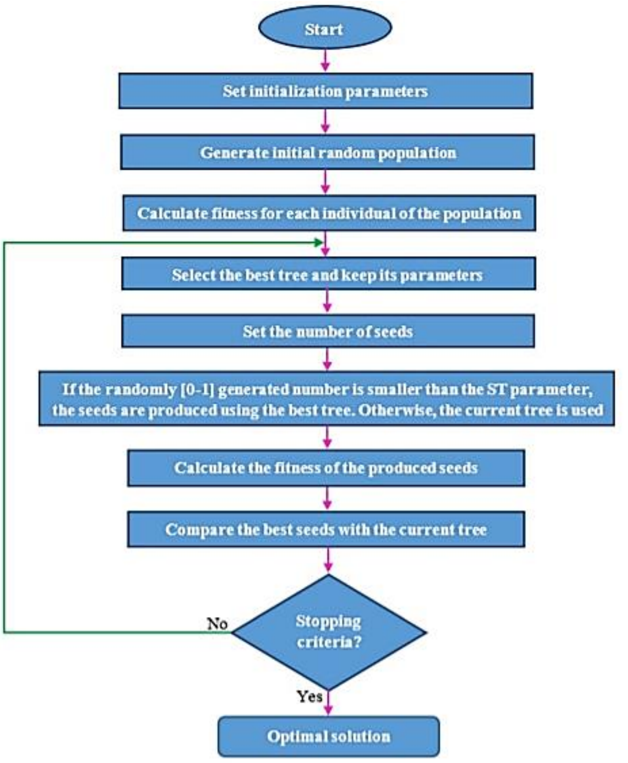

Start
Set initialization parameters
Generate initial random population
Calculate fitness for each individual of the population
Select the best tree and keep its parameters
Set the number of seeds
If the randomly [0-1] generated number is smaller than the ST parameter, the seeds are produced using the best tree. Otherwise, the current tree is used
Calculate the fitness of the produced seeds
Compare the best seeds with the current tree
No
Stopping criteria?
Yes
Optimal solution

Figure 3. Diagram of TSA

2.3.3. CSA Feature Selection

The Crow Search algorithm (CSA) is a meta-intuitive algorithm developed by Askarzadeh in 2016 [35]. The main idea underlying this algorithm emerged from crow behavior of hiding food. Crows have been deemed one of the smartest bird species as their brains are larger when compared to other birds. They have sophisticated facial recognition skills and warning systems that the use when encountering hostile situations. In addition, crows hide their food in safe places, sometimes for several months, and can always relocate their caches. Crows are considered thieves as they steal food from other birds. Thus, they benefit from predicting thief behaviours. They change locations when a theft is reported, allowing them to avoid future dangers.

The four main principles of CSA are defined as follows:

Crows live in flocks. Crows can remember where they hide their food caches. Crows watch out one another when they steal food.Crows are very wary of theft, possibly to preserve their caches. Algorithm provides the pseudo code of CSA.

```
Set the initial values of M, AP, f1 and tMax
Initialize the crow position y randomly
Evaluate the fitness function of each crow Fn(y)
Initialize the memory of search crow N
Set t=1. { Counter Initialization)
    repeat
        for(j=1 : j?M) do
            Randomly choose one of crows to follow z
            If $R_z$ ? $AP^{z,t}$ then
                $y^{j,t+1}= y^{j,t} + R_j * f1^{j,t} * ( N^{z,t} - y^{j,t} )$
            else
                $y^{j,t+1}$ = A random position of the sead
            end if
        end for
        Check the feasibility of $y^{j,t+1}$
        Evaluate the new position of crow $Fn(y^{j,t+1})$
        Update thee crow's memory $N^{j,t+1}$
        Set t=t+1 { Iteration counter increasing }
    Until (t < tMax ) { Termination criteria satisfied }
Produce the best solution N
```

2.3.4. SMA Feature Selection

A new meta-heuristic algorithm known as the Slime Mould Algorithm (SMA) has been proposed to solve continuous optimization problems [36]. SMA simulates the nature of slime mould using an oscillation mode to obtain the most convenient way to gather food with significant exploratory ability and tendency to concentrate [37]. The SMA model is divided into three stages: approach, wrapping, and food grabbing [38]. Algorithm provides the algorithm of the SMA model [39].

```
Input: N the number of solutions and total number of iterations ($t_{max}$)
Construct a random population (x)
T=1
while t<=$t_{max}$ do
    for each $X_i$ compute the fitness values ($F_i$)
    Update the value of the best solution $X_b$
    for i= 1 : N do
        Update $v_p$, $v_c$ and p using Eq. (3)
    t=t+1
Return the best solution $X_b$
```

2.3.5. ABC Feature Selection

Mimicking the foraging behaviour of a honey bee colony, the Artificial Bee Colony algorithm (ABC) was proposed by Karaboga in 2005 [40]. Bee food sources represent the optimization problems, and the amount of available nectar represents possible solutions and corresponding relevance values. ABC can be explained as follows for the optimization problem: Worker bees take advantage of previously discovered food sources and share information about the quality and location of the food sources with other bees by means of a waggle dance. The other bees in the hive decide upon the food source based on the information obtained from the worker bees. Scout bees are responsible for searching out new food sources based on an inner rule or a possible outer presumption [41-43]. Algoritm provides the pseudo code of the ABC algorithm [44].

```
Initialize population
Repeat
    Send the employed bees onto the food sources and evaluate
    their fitness (nectar amounts)
    for each employed bees do
        produce a new solution and determine its fitness
        apply greedy selection between new solution and current
        solution
    end
    evaluate the probability values of the food sources
    for each onlooker bees do
        select a food source depending on their fitness
        produce a new solutions and calculate its fitness
        apply greedy selection between new solution and current
        solution
end
abandon a position if the food source is exhausted by the bees
send the scout bees to the solution space for discovering new
food sources randomly for the abandoned positions
memorize the best food source found so far
until the stopping criteria are met;
```

2.4. Machine Learning Classifiers

In this section, features obtained from PSO, TSA, CSA, SMA, and ABC feature selection algorithms were classified by machine learning methods. Classification results were obtained by using classification algorithms and 10-fold cross validation methods. The SVM, DT, FFNN, RF, and kNN classification methods used are elaborated below.

2.4.1. The Decision Tree Method

The decision tree method (DT) has a simple, coherent structure, making it an appropriate method for the classification and prediction of problems. In decision trees, each node demonstrates a feature, each link (branch) represents a decision (rule), and each leaf shows a relevant result (categorical or continuing value) [45]. It is possible to syllogize by using decision tree data as they simulate human-level thinking. The main point is to create one tree for all data and demonstrate a single result on each leaf [46]. The decision tree method is a non-parametric supervised learning method used for classification and regression. DT is widely used in classification models due to its inexpensive setup, ease of results interpretation, ease of integration with database systems, and good reliability [47]. It has the advantages of visual, intuitive use and easy understanding [48].

2.4.2. The Random Forest Method

The random forest method (RF) is a set of decision trees and this classifier has become a popular option in machine learning processes. A random forest is created by using a large number of independent decision trees created by randomly selected variables. The algorithm itself creates the independent trees used. After the creation of the trees, they vote to find the most popular class [49]. The algorithm ensures the difference of all the trees in the forest. The randomness process is applied in two stages [50]. The first stage is to use different boot sample data to create each tree, and the second is to pick a random subset of estimators and interpret each tree node by dividing them into the best subset rather than overall interpretation [51]. There are two important reasons to apply the bootstrap stage: to increase the success of the classification accuracy using random features and to reduce generalization error [49]. The power of individual tree classifiers has importance in the classification process. At certain times, this algorithm works better than others such as support vector machines, neural networks, and discriminant analysis [52]. Random forests can also operate with other classifiers, even though they have been formed by decision trees. The optimum number of trees was defined as 50 in this study in order to obtain efficient results.

2.4.3. The Support Vector Machine Method

The Support Vector Machine method (SVM) is a classic machine learning technique that can assist solving large-scale data classification problems. It is especially useful in aiding multi-domain applications in a big data medium. The main idea of the SVM technique is to predict a model which requires the best hyperplane to separate the data. SVM has a solid theoretical basis and provides more accurate results when compared to other algorithms in many applications. As shown in Figure 4, SVMs use a maximal margin splitter [53]. This splitter represents the most remote control possible up to the sampling point. SVMs can also classify non-linear data by transferring data to a higher level via a method called kernel number [48].

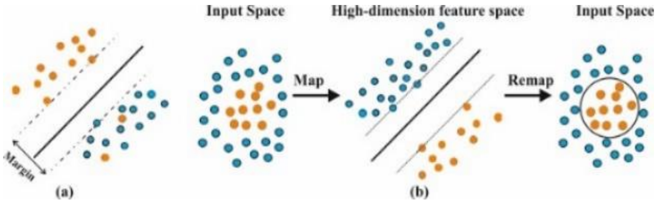


Figure 4. Model of SVM

2.4.4. The K-Nearest Neighbors Method

The K-Nearest Neighbors (kNN) method is a simple but effective method for classification [54]. This algorithm performs the classification process in accordance with the k value given as per the class of the nearest neighbor. In the kNN algorithm, vector classification is made using known vectors of the relevant class. The sample to be tested is processed separately with each sample in the training set. In order to determine the class of the sample to be tested, the nearest k sample is chosen in the training set. The sample to be tested belongs to this group, taking into consideration that which class has the most samples in the set of selected samples [48]. The kNN algorithm uses distance functions for the classification. In the present study, Euclid function was used, the distance significance for the Nearest Neighbor was equal, and the neighbor number was 5.

2.4.5. Feed Forward Neural Networks

ANNs have long been used for solving various complex engineering problems. ANN, is able to learn by using sample data. For this reason, it is a very useful model for revealing any correlation simulation that is difficult to describe mathematically or through physical models [55]. In most cases, neural networks can be used to make valid predictions. Feed forward neural networks, in particular, are used in classification problems encountered frequently in this regard [56]. In a FFNN, information is communicated from the inputs of the network to the outputs without feedback between the output layer and the input layer. Typically, neurons in each layer of the network receive their inputs only from the output signals of the previous layer [57].

2.5. K-Fold Cross Validation

Cross Validation (CV) is generally used in order to predict the generalization capability of a learning model [58] and is widely used in the evaluation of the performance of classification algorithms. Initially, a dataset is divided randomly into k-fold discrete pieces having approximately the same number of samples. Then, each layer tests the model induced by the other k-1 layers [59]. Cross validation is achieved by excluding one of them. At this point, one sample is kept as a verification/test sample at a time, and all remaining samples are used in training. Figure 5 provides the diagram of the K-Fold Cross Validation method.

2.6. Performance Evaluation

A confusion matrix was used in order to calculate the performance of the classification results obtained from the classification problems. Within the matrix there are four values: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Table1 demonstrates, using these four values, how many samples were correctly or incorrectly predicted for the binary classification.

Table 1. Binary Classification Confusion Matrix

		Predicted Condition	
	Total	**Diabetes**	**Non-Diabetes**
Actual Condition	**Diabetes**	TP	FN
	Non-Diabetes	FP	TN

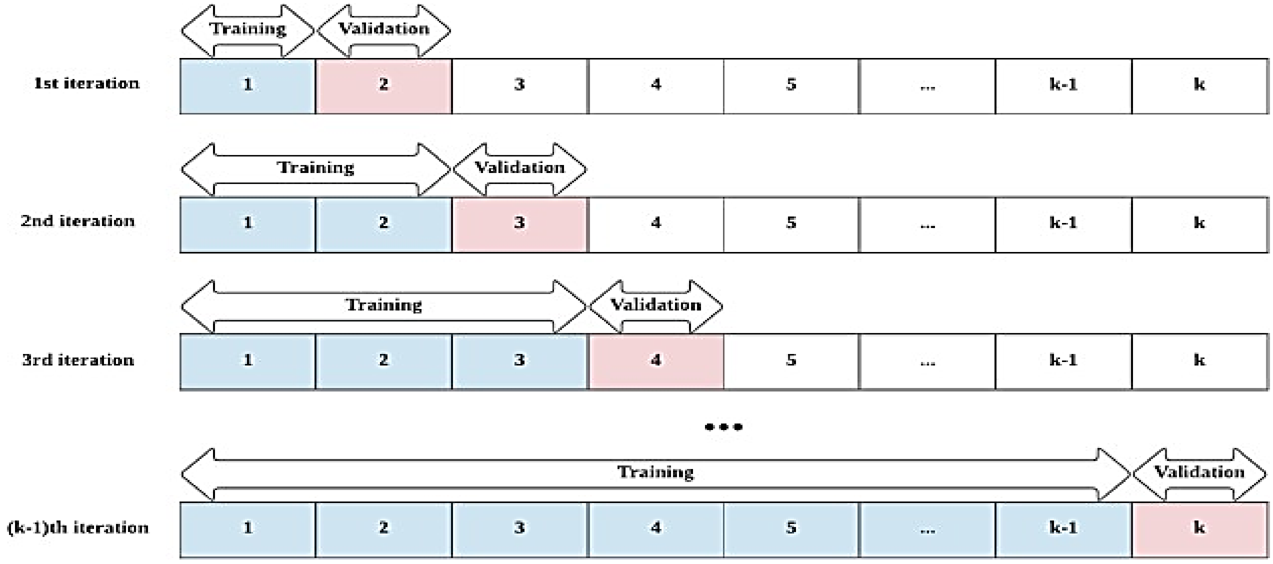


Figure 5. Diagram of K-Fold Cross-Validation

3. Results and Discussion

The optimum feature subsets were determined by using the encapsulating feature selection PSO, TSA, CSA, SMA and ABC algorithms in the present data set in order to evaluate the performance of the presented approach. Each algorithm specified in the study was run on the data set. The selected features are presented in Table 2.

Table 2. Features selected by each of the feature selection algorithms

Method	**Selected Features No**	**Number of Features**
PSO	2, 3, 4, 5, 8, 9, 11, 12, 14, 16	10
TSA	2, 3, 8, 9, 10, 12, 13, 14, 15	9
CSA	2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 16	13
SMA	2, 3, 4, 9, 15, 16	6
ABC	2, 3, 4, 7, 8, 10, 12, 15	8

The optimum features determined herein were classified using SVM, DT, FFNN, RF, and kNN machine learning methods and their performances were calculated thereof. The performance of the classifiers was measured with certain performance measurement techniques such as accuracy, precision, and recall data [60, 61]. Table 3 provides the results of the classification performance obtained from the k-fold cross-validation, where K=10.

Table 2 indicates that, of the wrapper-based feature selection methods, 10 were determined by the PSO method, 9 were determined by the TSA method, 13 were determined by the CSA method, 6 were determined by the SMA method, and 8 were determined by the ABC method. These respectively determined features were classified with the mentioned classification methods. In conclusion, methods displayed the best classification performances with success rates of PSO + SVM = 97.5, TSA + SVM = 96.15, CSA + FFNN = 99.04, SMA + FFNN = 94.23, and ABC + SVM = 96.73 respectively. Obtained accuracy values were compared with other studies in the literature and demonstrated in Table 4.

Table 3. Methods, Accuracy, Precision, Recall Values

Classifier Name	**Class (0/1)**	**n (truth)**	**N (classified)**	**Accuracy**	**Precision**	**Recall**
PSO Feature Selection						
DT	0	192	8	0.9558	0.96	0.9275
	1	15	305		0.9531	0.9744
RF	0	192	8	0.9577	0.96	0.932
	1	14	306		0.9563	0.9745
SVM	0	193	7	**0.975**	0.965	0.9699
	1	6	314		0.9813	0.9782
kNN	0	183	17	0.9519	0.915	0.9581
	1	8	312		0.975	0.9483
FFNN	0	190	7	0.9673	0.9645	0.95
	1	10	313		0.969	0.9781
TSA Feature Selection						
DT	0	190	10	0.9346	0.95	0.8879
	1	24	296		0.925	0.9673
RF	0	191	9	0.95	0.955	0.9183
	1	17	303		0.9469	0.9712
SVM	0	190	10	0.9558	0.95	0.936
	1	13	307		0.9594	0.9685
kNN	0	190	10	0.9577	0.95	0.9406
	1	12	308		0.9625	0.9686
FFNN	0	191	11	**0.9615**	0.9455	0.955
	1	9	309		0.9717	0.9656
CSA Feature Selection						
DT	0	187	13	0.9385	0.935	0.9078
	1	19	301		0.9406	0.9586
RF	0	192	8	0.9654	0.96	0.9505
	1	10	310		0.9688	0.9748
SVM	0	191	9	0.9808	0.955	0.9948
	1	1	319		0.9969	0.9726
kNN	0	193	7	0.9596	0.965	0.9324
	1	14	306		0.9563	0.9776
FFNN	0	197	2	**0.9904**	0.99	0.985
	1	3	318		0.9907	0.9938

Classifier Name	Class (0/1)	n (truth)	N (classified)	Accuracy	Precision	Recall
SMA Feature Selection						
DT	0	189	11	0.9308	0.945	0.8832
	1	25	295		0.9219	0.9641
RF	0	187	13	0.9308	0.935	0.8905
	1	23	297		0.9281	0.9581
SVM	0	190	10	0.9385	0.95	0.8962
	1	22	298		0.9313	0.9675
kNN	0	72	128	0.7192	0.36	0.8
	1	18	302		0.9438	0.7023
FFNN	0	192	22	**0.9423**	0.8972	0.96
	1	8	298		0.9739	0.9313
ABC Feature Selection						
DT	0	184	16	0.9385	0.92	0.92
	1	16	304		0.95	0.95
RF	0	193	7	0.9635	0.965	0.9415
	1	12	308		0.9625	0.9778
SVM	0	191	9	**0.9673**	0.955	0.9598
	1	8	312		0.975	0.972
kNN	0	190	10	0.9635	0.95	0.9548
	1	9	311		0.9719	0.9689
FFNN	0	190	10	0.9365	0.95	0.95
	1	10	310		0.9688	0.9688

Table 4. k-Fold Cross Validation Comparisons of Methods

Study	Method	Accuracy
(Das, et al. 2020)	Chi-square test (FS)+	
	DT	90.3
	kNN	91.3
	LR	85.57
	…	
	…	
	Max result	
	RFE-RF+ SVM	98.08
(Emon, et al. 2021)	LR	92
	DT	93
	Gaussian Process (GP)	88
	…	
	…	
	Max Result	
	RF	98
(Taser, 2021)	NBTree+AdaBoost	98.65
Recommended Approach	CSA (FS) + FFNN	99.04

4. Conclusion

Early diagnosis of diabetes mellitus is one of the most prominent topics in the literature today. Patient and non-patient data in public database clusters obtained from 520 samples collected via direct questionnaires from Sylhet Diabetes Hospital in Sylhet, Bangladesh was used in the present study [23]. Five feature selection algorithms were applied to this dataset within the scope of the present study. The features obtained from feature selection were classified using 5 classification methods from machine learning methods. An accuracy rate of 99.04% was achieved by means of the FFNN classification method using the features obtained from the CSA feature selection method, and Table 3 demonstrates the best classification accuracy result that can be determined with optimum properties obtained from feature selection algorithms.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Data Availability Statement

Data available in a publicly accessible repository (UCI Machine Learning Repository) that does not issue DOIs.

Ethical approval (and / or if people are involved) Ethical approval:

The study did not require ethical approval. Data available in a publicly accessible repository (UCI Machine Learning Repository) that does not issue DOIs.

Ethical confirmation:

This article does not contain any studies with animals by human participants or any authors.

References

[1] Kaddis, J.S., et al., Human Pancreatic Islets and Diabetes Research. JAMA, 2009. 301(15): p. 1580-1587.

[2] Sneha, N. and T. Gangil, Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big data, 2019. 6(1): p. 1-19.

[3] Zou, Q., et al., Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 2018. 9: p. 515.

[4] Organization, W.H., Diabetes fact sheet N 312. October 2013. Archived from the original on, 2013. 26.

[5] Organization, W.H., Guidelines for the prevention, management and care of diabetes mellitus. 2006.

[6] Saedi, E., et al., Diabetes mellitus and cognitive impairments. World journal of diabetes, 2016. 7(17): p. 412.

[7] Das, U., et al. Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers. in 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI). 2020. IEEE.

[8] Shetty, D., et al. Diabetes disease prediction using data mining. in 2017 international conference on innovations in information, embedded and communication systems (ICIIECS). 2017. IEEE.

[9] Singh, A. and R. Lakshmiganthan, Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. 2018.

[10] Ahmed, T.M., Using data mining to develop model for classifying diabetic patient control level based on historical medical records. Journal of Theoretical and Applied Information Technology, 2016. 87(2): p. 316.

[11] Singh, D., E.J. Leavline, and B.S. Baig, Diabetes prediction using medical data. Journal of Computational Intelligence in Bioinformatics, 2017. 10(1): p. 1-8.

[12] Azrar, A., et al., Data mining models comparison for diabetes prediction. Int. J. Adv. Comput. Sci. Appl., 2018. 9(8): p. 320-323.

[13] Wu, H., et al., Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 2018. 10: p. 100-107.

[14] Alam, T.M., et al., A model for early prediction of diabetes. Informatics in Medicine Unlocked, 2019. 16: p. 100204.

[15] Kavakiotis, I., et al., Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 2017. 15: p. 104-116.

[16] Islam, M.F., et al., Likelihood prediction of diabetes at early stage using data mining techniques, in Computer Vision and Machine Intelligence in Medical Image Analysis. 2020, Springer. p. 113-125.

[17] Rony, M.A.T., M.S. Satu, and M. Whaiduzzaman. Mining Significant Features of Diabetes through Employing Various Classification Methods. in 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). 2021. IEEE.

[18] Yue, C., et al. An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM. in 2008 International Symposium on Intelligent Information Technology Application Workshops. 2008. IEEE.

[19] Fiarni, C., E.M. Sipayung, and S. Maemunah, Analysis and prediction of diabetes complication disease using data mining algorithm. Procedia Computer Science, 2019. 161: p. 449-457.

[20] Ahmed, T.M., Developing a predicted model for diabetes type 2 treatment plans by using data mining. Journal of Theoretical and Applied Information Technology, 2016. 90(2): p. 181.

[21] Emon, M.U., et al. Primary Stage of Diabetes Prediction using Machine Learning Approaches. in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). 2021. IEEE.

[22] Taser, P.Y. Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction. in Multidisciplinary Digital Publishing Institute Proceedings. 2021.

[23] URL1.Https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. 2020.

[24] Pudil, P., J. Novovičová, and J. Kittler, Floating search methods in feature selection. Pattern recognition letters, 1994. 15(11): p. 1119-1125.

[25] Dash, M. and H. Liu, Feature selection for classification. Intelligent data analysis, 1997. 1(1-4): p. 131-156.

[26] Nguyen, H.B., et al. Filter based backward elimination in wrapper based PSO for feature selection in classification. in 2014 IEEE congress on evolutionary computation (CEC). 2014. IEEE.

[27] Tran, B., B. Xue, and M. Zhang, A new representation in PSO for discretization-based feature selection. IEEE Transactions on Cybernetics, 2017. 48(6): p. 1733-1746.

[28] Amoozegar, M. and B. Minaei-Bidgoli, Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism. Expert Systems with Applications, 2018. 113: p. 499-514.

[29] Almomani, O., A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms. Symmetry, 2020. 12(6): p. 1046.

[30] Nemati, S. and M.E. Basiri. Particle swarm optimization for feature selection in speaker verification. in European Conference on the Applications of Evolutionary Computation. 2010. Springer.

[31] Kiran, M.S., TSA: Tree-seed algorithm for continuous optimization. Expert Systems with Applications, 2015. 42(19): p. 6686-6698.

[32] Babalik, A., A.C. Cinar, and M.S. Kiran, A modification of tree-seed algorithm using Deb's rules for constrained optimization. Applied Soft Computing, 2018. 63: p. 289-305.

[33] Chen, F., et al. A feature selection approach for network intrusion detection based on tree-seed algorithm and k-nearest neighbor. in 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS). 2018. IEEE.

[34] Beşkirli, A., D. Özdemir, and H. Temurtaş, A comparison of modified tree–seed algorithm for high-dimensional numerical functions. Neural Computing and Applications, 2019: p. 1-35.

[35] Askarzadeh, A., A novel metaheuristic method for solving constrained engineering optimization problems: crow search algorithm. Computers & Structures, 2016. 169: p. 1-12.

[36] Cheng, R., et al. Benchmark functions for CEC'2017 competition on evolutionary many-objective optimization. in Proc. IEEE Congr. Evol. Comput. 2017.

[37] Abdel-Basset, M., et al., An efficient binary slime mould algorithm integrated with a novel attacking-feeding strategy for feature selection. Computers & Industrial Engineering, 2021. 153: p. 107078.

[38] Li, S., et al., Slime mould algorithm: A new method for stochastic optimization. Future Generation Computer Systems, 2020. 111: p. 300-323.

[39] Ewees, A.A., et al., Improved Slime Mould Algorithm based on Firefly Algorithm for feature selection: A case study on QSAR model. Engineering with Computers, 2021: p. 1-15.

[40] Karaboga, D., An idea based on honey bee swarm for numerical optimization. 2005, Citeseer.

[41] Karaboga, D. and B. Basturk, A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of global optimization, 2007. 39(3): p. 459-471.

[42] Das, S., S. Biswas, and S. Kundu, Synergizing fitness learning with proximity-based food source selection in artificial bee colony algorithm for numerical optimization. Applied Soft Computing, 2013. 13(12): p. 4676-4694.

[43] Hancer, E., et al., A binary ABC algorithm based on advanced similarity scheme for feature selection. Applied Soft Computing, 2015. 36: p. 334-348.

[44] Moosa, J.M., et al., Gene selection for cancer classification with the help of bees. BMC medical genomics, 2016. 9(2): p. 135-165.

[45] Jadhav, S.D. and H. Channe, Efficient recommendation system using decision tree classifier and collaborative filtering. Int. Res. J. Eng. Technol, 2016. 3(8): p. 2114-2118.

[46] Patel, H.H. and P. Prajapati, Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering, 2018. 6(10): p. 74-78.

[47] Kotsiantis, S.B., I. Zaharakis, and P. Pintelas, Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 2007. 160(1): p. 3-24.

[48] Koklu, M. and I.A. Ozkan, Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture, 2020. 174: p. 105507.

[49] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

[50] Gokgoz, E. and A. Subasi, Comparison of decision tree algorithms for EMG signal classification using DWT. Biomedical Signal Processing and Control, 2015. 18: p. 138-144.

[51] Yi, Z. and J. Pan. Application of random forest to stellar spectral

classification. in 2010 3rd International Congress on Image and Signal Processing. 2010. IEEE.

[52] Liaw, A. and M. Wiener, Classification and regression by randomForest. R news, 2002. 2(3): p. 18-22.

[53] Nakano, T., et al., Gaits classification of normal vs. patients by wireless gait sensor and Support Vector Machine (SVM) classifier. International Journal of Software Innovation (IJSI), 2017. 5(1): p. 17-29.

[54] Guo, G., et al. KNN model-based approach in classification. in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". 2003. Springer.

[55] Khudhair, A. and N.A. Talib. Neural Network Analysis For Sliding Wear of 13% Cr Steel Coatings by Electric Arc Spraying. in Diyala Journal of Engineering Sciences-First Engineering Scientific Conference, College of Engineering–University of Diyal. 2010.

[56] Mehrotra, K., C.K. Mohan, and S. Ranka, Elements of artificial neural networks. 1997: MIT press.

[57] Sreekanth, P., et al., Comparison of FFNN and ANFIS models for estimating groundwater level. Environmental Earth Sciences, 2011. 62(6): p. 1301-1310.

[58] Jiang, G. and W. Wang, Error estimation based on variance analysis of k-fold cross-validation. Pattern Recognition, 2017. 69: p. 94-106.

[59] Wong, T.-T. and P.-Y. Yeh, Reliable accuracy estimates from k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering, 2019. 32(8): p. 1586-1594.

[60] OZKAN, I.A. and M. KOKLU, Skin lesion classification using machine learning algorithms. International Journal of Intelligent Systems and Applications in Engineering, 2017. 5(4): p. 285-289.

[61] Cinar, I. and M. Koklu, Classification of Rice Varieties Using Artificial Intelligence Methods. International Journal of Intelligent Systems and Applications in Engineering, 2019. 7(3): p. 188-194.