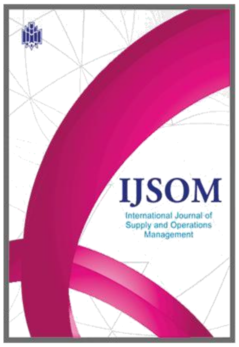
International Journal of Supply and Operations Management

IJSOM

2024, Volume 11, Issue 4, pp. 462-482
ISSN-Print: 2383-1359
ISSN-Online: 2383-2525
www.ijsom.com



Assessment of a Hybrid Machine Learning Algorithm in Healthcare Management for Predicting Diabetes Disease

Azin Nodoust [a*] and Ali Rajabzadeh Ghatari [a]

[a] *Department of Industrial Management, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran*

Abstract

Diabetes Mellitus is one of the most chronic diseases in all over the world. Every year, many people die due to this disease in all countries. Therefore, identifying early detection methods for this disease can reduce its mortality. Today, many diseases can be diagnosed and prevented from progressing by using data mining techniques and machine learning algorithms. In this paper, diabetes prediction has been aimed by comparing the efficiency of several classical machine-learning techniques. For this reason, for the sake of diabetes prediction algorithms such as Naïve Bayes, Logistic Regression (LR), Multi-Layer Perceptron (MLP), Sequential Minimal Optimization (SMO), J48, Random Forest (RF), Regression Tree (RT) algorithms and a new hybrid algorithm based on Multi-Verse Optimizer (MVO) and Multi-Layer Perceptron (MLP) algorithms are employed for this evaluation based on Accuracy (ACC) Indicator and Area under Curve (AUC) criteria. Numerous and diverse methods and algorithms have been used to predict diabetes. Each of these algorithms has been effective in predicting diabetes with a different level of accuracy. Our goal in this research is to introduce a new combined algorithm that has the highest level of accuracy in predicting diabetes compared to the old frequent algorithms so that it can help people in the timely treatment of this disease. In the structure of the MLP algorithm, the backpropagation algorithm is used for training. This article uses the MVO algorithm to train the MLP instead of the backpropagation algorithm, which built the hybrid algorithm called MVO-MLP. The accuracy results and the area under the ROC diagram Indicated that the proposed hybrid algorithm increases the accuracy by 107% compared to the MLP algorithm with the default structure. The outcomes of the accuracy of the new model are also higher than other algorithms used in this article

Keywords: Diabetes Mellitus; Machine Learning Algorithm; Data Mining; Accuracy; Area under Curve; Multi-Verse Optimizer; Multi-Layer Perceptron.

1. Introduction

Diabetes Mellitus (DM) is characterized as a collection of metabolic clutters primarily arising from the abundance of glucose inside the bloodstream (Shaw et al. 2010). In other words, diabetes is a system of immunity infection where in the significant cells that create insulin for retaining glucose are destroyed which are required to create energy within the body (Bellamy et al. 2009). The pancreas releases insulin that makes a difference in humans to urge energy but if an individual has diabetes the human body isn't able to produce adequate insulin that utilizes the insulin delivered (Olokoba et al. 2012 in positive diabetes individuals, either

*Corresponding author email address: n.azin@modares.ac.ir

DOI: 10.22034/ijsom.2024.110346.3067

the pancreas does not make enough insulin, or bodies without insulin, on the other hand, the cells are not able to respond appropriately to the insulin which is generated consequently, substantial volume of glucose enters the bloodstream, which is poured into the urine by the keys and leaves the body. Therefore, the body no longer has its principal origin of fuel, which accommodates a considerable volume of glucose.

DM is one of the most frequent internal secretion disarranges, influencing over 200 million people around the world (Cox & Edelman, 2009). This is the fifth foremost foundation of passing away in ladies and the eighth focal basis of passing away for both genders in 2012.

These days, numerous individuals are enduring from DM. In all directions, 425 million individuals endure diabetes following 2017 insights. Around 2-5 million ailing annually pass away on account of this crisis. It has been pronounced that by 2045 this will upheave to 629 million (Kalyankar et al. 2017).

There are two important types of diabetes: Type 1 diabetes called dependence insulin diabetes which most appear more regularly at an early age. In Type 1, the pancreas with immunizer is assaulted by the body, after that it annihilates inside parts of the body and prevents insulin production. Type 2 is additionally called adult-onset diabetes or non-insulin-dependent. In most cases, it is more tolerant than Type 1, but it is still exceptionally destructive (Himsworth & Kerr, 1939).

Due to the incurable complications of diabetes, the use of tools to predict the disease is inevitable. Machine learning techniques can be a conductive implemented tool for predicting the disease. Without a doubt, subsequently, machine learning and data mining procedures in DM are of incredible concern when it considers of determination, administration, and other related clinical ingredients (Kavakiotis et al., 2017).

Machine learning is the logical area of managing how machines learn from encounters. For numerous researchers, the expression “Machine Learning” is indistinguishable from the turn of phrase “Artificial Intelligence”, given that the plausibility of learning is the most distinguishing of an establishment called intelligent in the extensive aspect of the word. The dominant intention of machine learning is the development of the frameworks of computers that can adjust and learn from their involvement (Wilson & Keil, 1999).

There are four headings of machine learning models: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinfcmant learning. In this paper supervised algorithms have been used in comparison with the proposed algorithm.

A. Supervised Learning

A supervised learning technique accommodates a collection of input data and a group of outputs and constructs a model to create reasonable predictions for the reaction to the unused dataset. Decision Trees, Artificial Neural Network Bayesian Method, Ensemble techniques, and Instance-based methods, are examples of supervised learning algorithms (Nithya & Ilango, 2017).

B. Unsupervised Learning

Unsupervised Learning methodology is used to develop descriptive models. Descriptive models contain of known inputs and unknown outputs. unsupervised learning is generally used on value-based data. k-Medians and K-means clustering are examples of unsupervised learning algorithms (Nithya & Ilango, 2017).

C. Semi-Supervised Learning

either marked and unmarked data are used by Semi-Supervised methodology. The semi-supervised methodology contains Regression techniques and Classification. A certain amount of the algorithms of Regression techniques in Semi-Supervised learning are Logistic Regression and Linear Regression (Nithya & Ilango, 2017).

D. Reinforcement Learning

The difference between Reinforcement Learning and Supervised Learning is that reinforcement learning does not need inputs and outputs with labels. but the function of it is to find an adjustment between investigation (of the unfamiliar region) and misuse (of current information) to maximize the long-term compensation. (Kaelbling et al.,1996)

Therefore, in this paper, a hybrid framework of machine learning called MLP training by MVO (MVO-MLP) has been proposed. This research aims to compare the performance of the prediction of diabetes disease, with the proposed algorithm and other classification and regression methods of supervised machine learning algorithms such as Naïve Bayes, Random Tree, Multi-layer perceptron, J48, Random Forest, Logistic Regression, Sequential Minimal Optimization and Support Vector Machines (SVM). The results indicate that the proposed model has the best sequel in diagnosing Diabetes Mellitus.

2. Related Work

In the literature, many researchers have investigated different methods for timely diagnosis of diabetes before it endangers people's lives.
Different studies have been conducted to predict the side effects of diabetes by applying diverse approaches such as machine learning and data mining (Kavakiotis et al. 2017). The related work and the challenges are in the table 1 below:

Table 1. Literature Review

Authors	Used Technologies	Key Findings	Challenges
Gowthami et al. (2024)	They utilized machine learning algorithms such as Logistic Regression, K- Nearest Neighbors, Decision Trees, Random Forest, and Support Vector Machines.	The results demonstrate that the Random Forest algorithm achieves an impressive accuracy of 98 %.	Using old and repetitive algorithms.
Gangani et al. (2024)	They used (Decision Tree (DT), K-nearest Neighbor (KNN), Support Vector Classification (SVC), and Extreme Gradient Boosting (XGB).	The Gradient Boosting (XGB) achieved the highest accuracy.	There should be more than 3 classification algorithms.
Samsel et al. (2024)	They used Logistic Regression, Random Forest, AdaBoost, XGBoost, Naive Bayes, and Artificial Neural Networks for diagnosing diabetes.	XGBoost emerged as the most effective model with an AUC of 0.70	There is just AUC as the criteria and there is a lack of using ACC as an important method.
Khaleel and Al-Bakry (2023)	They provided one superior model to predict if the result of the diabetes test on each person was positive or negative. They used different methods like precision, recall, and F1-measure for this research.	Proposed algorithm was Logistic Regression which obtained 94% of accuracy compared to K-nearest Neighbor and Naïve Bayes algorithms.	The introduced algorithm is Repetitious and old.
Chang et al. (2023)	They analyzed the J48 decision tree, Random forest models, and Naïve Bayes for both testing and training the PIMA dataset.	They found Naïve Bayes was a better evaluator if kinds of exact feature selection were used and Random Forest was the best if the features were numerous.	The algorithms are too old and there should be more classification methods.
Ahmed et al. (2023)	They Indicated one framework called the Fused ML model in diabetes diagnosis compared to Artificial Neural Network (ANN) models plus the Support Vector Machine (SVM) technique.	The new model achieved an accuracy of 94.87 %.	The number of algorithms are not sufficient and there should be more algorithms to prove the result.
Kangra and Singh (2023)	They used different kinds of machine learning such as Naïve Bayes, K nearest neighbor, random forest, logistic regression, and decision tree algorithms. Finally, support vector machines on two types of data such as German diabetes data and Indian diabetes datasets have been considered.	The support vector technique was introduced which achieved the best result.	The German data set is not well known to count on and there is no AUC as the second criteria. The algorithms are also old and there is no hybrid algorithm.
Raja Krishnamoorthi et al. (2022)	They used support vector machine (SVM) learning models and decision tree (DT)-based random forest (RF) and one proposed logistic regression algorithm for diabetes prediction	The proposed logistic regression algorithm was introduced as the best accuracy criterion.	The algorithms are old and there is no hybrid model

Table 1. Literature Review (*Continued*)

Lu et al. (2022)	The risk of chronic disease was measured using a Support Vector Machine, Logistic Regression technique, Naïve Bayes, K-nearest Neighbors, Random Forest, Decision Tree, XGBoost, and finally Artificial Neural Network.	They proposed The Random Forest as the impressive model with a variable rate of AUC between 0.79 to 0.91 compared to others.	The introduced random forest algorithm is old and there should be hybrid algorithms to innovate the work
El Massari et al. (2022)	They utilized ACC, Recall, F-Measure, and Precision as conductive metrics to perform the risk of diabetes disease.	They introduced ontology classifiers and SVM as better algorithms compared to Decision Trees, KNN, Logistic regression, Naive Bayes, and also ANN as analyzers.	The introduced algorithm is old and there should be a hybrid algorithm as a new one
Rawat et al. (2022)	They used Adaboost, Naïve Bayes (NB), Neural Network, Support Vector Machine (SVM), and also K-nearest neighbor (KNN)	A Neural Network has been investigated as a successful algorithm to forecast the possibility of positive or negative diabetes diseases.	There is a lack of other classification methods.
Victor et al. (2022)	They utilized a Decision Tree, Random forest, K-nearest Neighbor's algorithm, and Naïve Bayes algorithms, besides Logistic Regression, to predict diabetes disease. The authors used ACC as the criteria for prediction.	They proposed Random Forrest as the best one compared to the others.	The authors did not use AUC besides ACC to make sure of the result. The algorithms are old and it is better to innovate the paper with the hybrid algorithm.
Theerthagiri et al. (2022)	They compared the MSE and Accuracy of some algorithms like, Naive Bayes, Multilayer Perceptron, Radial Basis Function, Extra Trees, K-nearest Neighbor, and Decision Trees to calculate the possibility of having diabetes in the future.	They introduced MLP as the best criteria with the lowest level of MSE and the highest level of accuracy.	There is a lack of using AUC besides ACC as criteria and the algorithms are old.
Kumari et al. (2021)	Another machine learning algorithm named Soft Voting Classifier has been proposed by authors in comparison to the other classifiers like Naïve Bayes, Logistic Regression, AdaBoost, Support Vector Machine, GradientBoost, Random Forest, XGBoost, and also CatBoost for diagnosing diabetes.	The suggested model achieved a remarkable percentage of accuracy among others.	The accuracy is not enough for this paper as criteria.
Ganie and Malik (2021)	Authors used the K-fold cross-validation and feature engineering method and some evaluators such as misclassification rate (MCR), accuracy rate, F1-score precision, specificity, recall, and also receiver operating characteristic (ROC) curve to introduce one algorithm between Bagging, Boosting, and Voting.	They proposed a bagged decision tree as the selected one.	Training and testing methods are lost in this paper.
Islam et al. (2020)	Some other machine learning algorithms such as SVM, Linear Discriminant Analysis, and Bagged CART have been investigated.	They found 11 important risk factors of diabetes disease and Bagged CART achieved the highest accuracy.	The number of algorithms is not sufficient.
Lukmanto et al. (2019)	They utilized the F-score method with feature selection and a hybrid model called FUZZY SVM for the classification of diabetes disease.	The result indicated 89.02% of accuracy.	The f-score is not the best method to measure accuracy. in this paper, there is a lack of using ACC and AUC methods.

Table 1. Literature Review (*Continued*)

Alam et al. (2019)	They used the PCA method for feature selection and ANN, RF as classifiers, and K-means as clustering algorithms.	They found Body Mass Index (BMI) and glucose level as the best features and also ANN algorithms as the best classifier with 75.7% level of accuracy.	The number of classification and clustering algorithms should be more.
Zou et al. (2018)	Principal Component Analysis (PCA) beside Maximum Relevance-Minimum Redundancy (MRMR) was used with feature selection criteria and J48, Artificial Neural Networks (ANN), beside Random Forest with a method of classification.	The accuracy of using MRMR was better than the PCA method.	There is a lack of using some combined algorithms and more algorithms.
Fatima and Pasha (2017)	They used J48, AdaBoost, and bagging algorithms of machine learning to propose the best algorithm for predicting diabetes or non-diabetes disease.	AdaBoost showed a better outcome compared to others.	There is a lack of using some combined algorithm.
Shetty et al. (2017)	They presented the Naïve Bayes beside K-Nearest Neighbors (KNN) techniques to forecast diabetes disease, using the positive diabetes people information and records as inputs to get the result in the form of diabetes or not.	They found by using Naïve Bayes and K-Nearest Neighbors the result of accuracy will be improved compared to other methods of older papers.	The lack of other classification methods.
Hina et al. (2017)	Random Forest (RF), Multi-Layer Perceptron (MLP), Logistic Regression (LR), Naïve BayesJ48, and Zero R, have been applied by the authors.	The result proposed MLP, as the best algorithm in terms of efficiency and accuracy compared to the rest.	The algorithms used are so old and there is a lack of using some combined algorithm.
Ahmed, T.M. (2016)	The Naïve Bays, Logistic, and J48 were used to predict diabetes.	The logistic algorithm has been selected as the best one with a high accuracy rate of 74.8%.	The lack of other classification methods.
Butwall, M. and S. Kumar(2015)	They used the machine learning algorithm called Random Forest in their proposed model to diagnose diabetes disease.	Random Forest Classifier-based approach outperforms better with an accuracy of 99.7%.	h particular lifestyle parameters, including physical activity and emotional states, the only parameters and they did not use other classifier.
Iyer et al. (2015)	They utilized Decision tree algorithms beside Naïve Bayes to discover the covered-up patterns in the datasets of diabetes.	They found by using Naïve Bayes and Decision tree the result of accuracy will be improved compared to other methods of older papers.	The lack of other classification methods.
Marcano-Cedeño et al. (2011)	They extended a hybrid model called (AMMLP) using Multi-Layer Perceptron (MLP) trained by Artificial Meta Plasticity (AM), to make the prediction model to diagnose diabetes disease. They used the PIMA dataset and WEKA tool for the prediction.	The conclusion of the proposed framework indicated 89.93% accuracy for this hybrid model.	Using Python was better for hybrid algorithms and this paper just used the Weka tool.
Ahmed et al. (2011)	They evaluated the accuracy of one famous algorithm called Multi-Layer Perceptron (MLP), in opposition to J48 and ID3 algorithms.	The J48 algorithm achieved higher accuracy and performance in predicting diabetes disease.	The algorithms used are so old and there is a lack of using some combined algorithm.
Khan and Mohamudally (2011)	They utilized different machine learning classifiers and clusters such, as Neural Networks, C4.5 Decision Tree, and K-Means to predict diabetes disease.	They found the decision tree the best algorithm with the highest accuracy result.	The lack of other classification methods.

Table 1. Literature Review (*Continued*)

Patil et al. (2010)	They offered a compound model of machine learning to forecast diabetes disease. In their hybrid model called HPM. they used K-means clustering to validate of class label and also C4.5, to build the final model.	The proposed model illustrated 92.33% of accuracy.	The algorithms used to combine are old.
Polat and Güneş (2007)	They used a combination of Principal Component Analysis (PCA) and Neuro-Fuzzy conjecture tools to predict diabetes.	They found their hybrid algorithm with an accuracy of 89.4% algorithms on literature compared to other.	They didn't use other classification algorithms on their work They just brought other related work.

By reviewing the history of literature, we realize that in health science, the use of machine learning algorithms for the prediction of various diseases, especially the diagnosis of diabetes, has become old and repetitive. Classical algorithms that are used in the prediction of diabetes have limited accuracy in predicting this disease, and it is necessary to use alternative algorithms with higher accuracy in diagnosing diabetes.it is better to go towards the use of new combined algorithms that gives a higher level of accuracy for diagnosing diabetes Therefore, the combined algorithm that has been used in this study is MVO-MLP which presents more level of accuracy for the prediction of diabetes that is higher than other algorithms's results. so it can be said that this novel combined algorithm can be used as a conductive way to prevent diabetes with a higher level of convenience which is effective in timely treatment

In this paper, the Naïve Bayes, Multi-Layer Perceptron (MLP), J48, Sequential Minimal Optimization (SMO), Logistic Regression (LR), Regression Tree (RT), and Random Forest (RF) algorithms have been investigated to predict diabetes by comparing the performance of them. Moreover, a new hybrid algorithm based on Multi-Verse Optimizer (MVO) and also Multi-Layer Perceptron (MLP) algorithms are presented for this evaluation based on Accuracy (ACC) metric and Area under Curve (AUC) criteria. The results show that the new hybrid algorithm predicts diabetes more accurately than the others used for this prediction.

In Section 2 the Multi-Verse Optimizer (MVO) and Multi-Layer Perceptron (MLP) algorithms which the new hybrid algorithm is based on, and the reason for making this new hybrid algorithm is described Section 3 represents the data preparation, the Materials, and the evaluation method used. In section 4 the results are represented, and finally, the conclusion and future of our work are presented.

3. Methodology

For presenting a new hybrid model, we have used two types of algorithms. The algorithms are MVO (Multi-Verse Optimizer) and MLP (Multi-Layer Perceptron. the reason of using MVO as a trainer for MLP instead of a backpropagation algorithm and making this hybrid algorithm is to find whether the newly developed algorithms could be better trainers and give better results of accuracy in predicting diabetes diseases. Each of these algorithms is described below to show their independent function.

1.3. Multi-Verse Optimizer (MVO)

The MVO algorithm is an algorithm based on the Big Bang theory which was inspired by nature. According to this theory, large explosions in the universe cause parallel universes to form. In this theory, the three most important principles are black holes, white, and wormholes, which are used as basic notions in the MVO algorithm and are modeled using mathematical concepts. In the MVO algorithm, parallel worlds and objects within them, are used as variables and solutions. The transfer of objects and the exchange between these worlds take place constantly. So in the MVO algorithm, the purpose is looking to find the best world for moving the objects. (Mirjalili et al., 2015).

Figure 1 indicates a Multi-Verse Optimizer (MVO).

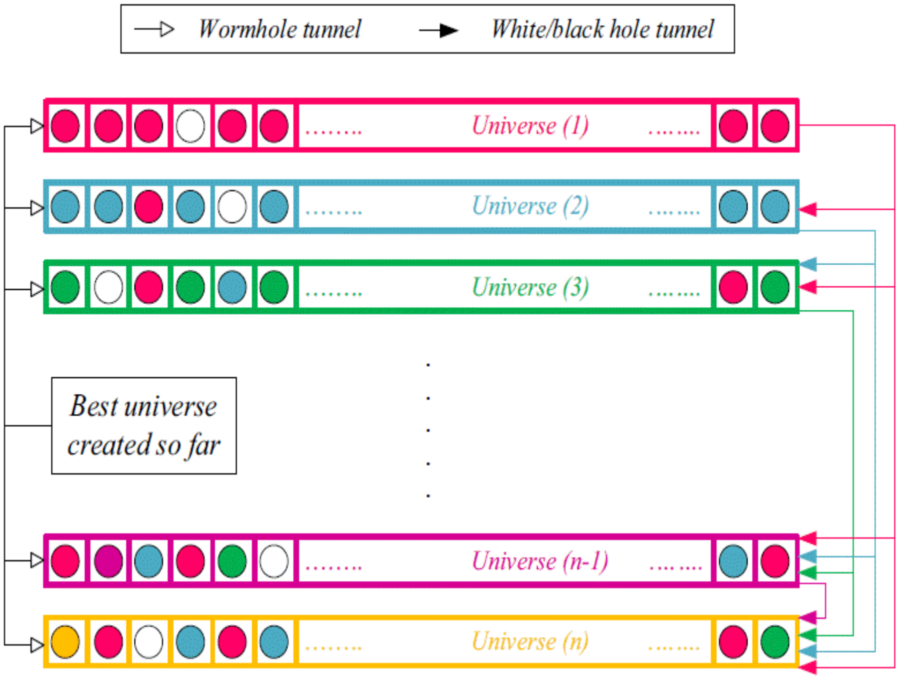


Figure 1. Multi-Verse Optimizer (MVO)

Finding the best world is done using the Roulette wheel selection (RWS) algorithm, the steps of which are given below:

If the consideration is on U as a set consisting of n universes and consider d as the number of variables in each universe, the set U is as follows:

$$U = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}$$

In order to achieve white holes, the worlds based on the inflation rate must be selected, which is calculated as follows:

$$x_{ij} = \begin{cases} x_{kj} & r1 < NI(U_i) \\ x_{ij} & r1 \geq NI(U_i) \end{cases}$$

Therefore, U_i represents each of the worlds and N_i represents the inflation rate of each world, X_{ij} reveals the jth variable of the ith universe, and r_1 randomly is chosen between [1]. Additionally, X_{kj} presents the jth variable of the kth universe.

Assuming that wormholes cause objects to move between one world and the best of the world, the mechanism expressed is formulated as follows:

$$x_{ij} = \begin{cases} \begin{cases} x_j + TDR \times \left(\left(ub_j - lb_j \right) \times r_4 + lb_j \right) & r3 < 0.5 \\ x_j - TDR \times \left(\left(ub_j - lb_j \right) \times r_4 + lb_j \right) & r3 \geq 0.5 \end{cases} & r2 < WEP \\ x_{ij} & r2 \geq WEP \end{cases}$$

Where WEP is the probability of an existing wormhole and TDR is the rate of distance traveled between the universe. The formulate of calculating WEP and TDR is represented as follows:

$$WEP = W_{min} + l \times \left(\frac{W_{Max} - W_{min}}{L}\right)$$

Where W_{min} and W_{Max} are equal to 0.2 and 1 respectively. Moreover, L is the maximum number of replications.

$$TDR = 1 - \frac{l^{1/p}}{L^{1/p}}$$

Where P equals to 6 as a constant number for accelerating the exploitation accuracy.

2.3. Multi-Layer Perceptron (MLP) Neural Network

An Artificial Neural Network (ANN) is a type of machine learning technique and mathematical structure, which is designed to simulate the brain neurons for processing the data. (Hinton 1992; Jensen 1994).

A kind of the most introduced type of ANN is Multi-Layer Perceptron (MLP). The basic structure of MLP is made of three layers such that: (1) the input layer that takes the input data (2) the hidden layer that is used for feature processing and (3) the output layer which indicates the final results. Figure 2 indicates a Multi-Layer Perceptron (MLP) Neural Network.

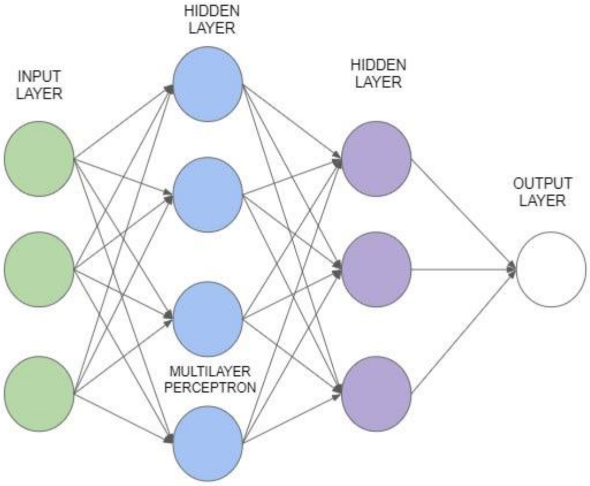


Figure 2. Multi-Layer Perceptron (MLP) Neural Network

The structure of MLP is based on the Trial-and-Error technique. In the MLP model, the input layers (x_1 to x_j) are multiplied by the weight assigned to each of them (W_{in} to W_{jk}) and then the net input (Net_i) is calculated, by adding a threshold (b_t) to the input. The exact value of Net_i is always higher than zero. During the process, the weights assigned to each input are continuously optimized by the learning operation.

3.3. Proposed Work

In general, utilizing a backpropagation algorithm to train MLP is implemented. In this paper, a new hybrid model called MVO-MLP has been performed, which uses the MVO algorithm for training the MLP mechanism. The motivation of this study is to compare the function of the presented algorithm with the most applicable machine learning algorithms such as Random Forest (FR), Sequential Minimal Optimization (SMO), Logistic Regression (LR), Naïve Bayes, Multi-Layer Perceptron (MLP), J48, and Regression Tree (RT) algorithms to forecast Diabetes Mellitus.

4. Material and Methods

The material and methods used in this paper are Dataset Description which **is** the source of the dataset and their type, Data Generation which shows the levels of preparing and cleaning data to use, the software and tools for running the data. and Data Evaluation which shows the methods and metrics of evaluating data. the methods in this paper are the kind of preparing data to use as inputs and the metrics for evaluating the performance of algorithms. the methods and materials are described below

4.1. Dataset Description

In this study, the Pima Indian Dataset (PID) of diabetes disease has been used, which is collected and sourced from the Kaggle Repository and contains 763 rows which indicate the number of patients and 10 columns of features. The last column indicates whether people have positive diabetes or negative. In this data set, a total of 268 people have positive diabetes, and 500 people have negative one. The description of this dataset features 20 samples of these people are brought in Table 2.

Table 2. The description of dataset features for 20 samples

Number of pregnant	Plasma glucose concentration	Diastolic blood pressure	Triceps skin fold thickness	2-Hours serum insulin	Body mass index	Diabetes pedigree function	Age	Class
6	148	72	35	0	33.6	0.627	50	0
1	85	66	29	0	26.6	0.351	31	1
8	183	64	0	0	23.3	0.672	32	0
1	89	66	23	94	28.1	0.167	21	1
0	137	40	35	168	43.1	2.288	33	0
5	116	74	0	0	25.6	0.201	30	1
3	78	50	32	88	31	0.248	26	0
10	115	0	0	0	35.3	0.134	29	1
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	0
4	110	92	0	0	37.6	0.191	30	1
10	168	74	0	0	38	0.537	34	0
10	139	80	0	0	27.1	1.441	57	1
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	0
1	103	30	38	83	43.3	0.183	33	1
1	115	70	30	96	34.6	0.529	32	0

4.2. Data Generation

One of the most important ways to make useful data is data preparation. In this paper, three levels of data preparation techniques have been used. At first, data normalization has been utilized in order to replace the values in ranges of [0,1], Then the label encoding has been used to replace the string rates using 0 and 1in the output, finding the person with positive or negative diabetes. At the end of preparation, the missing rates were restored by the median.

4.3. Tools

Weka and Python were selected for running the experiments. The reason for using Weka for this paper is that it is implemented for running the classification and supervised algorithms and in this paper there are classification algorithms that were named before. for running and coding the hybrid algorithm MVO-MLP python has been used to give the best result. On the other hand, for training MLP with MVO, the python gave the best results. The number of running each kind of algorithm is 10 and the average of them is calculated. 30% of the Pima data set were used for testing and 70% were used to train.

Testing and Training

Training is a process of fitting or classifying the parameters like weights. The training data set is the best data for the supervised algorithm to make the prediction models to make a training model to generate a good predictive model. So it is better to use most of the datasets to train with supervised algorithms and give the best result.(Larose, D. T 2014)

Testing data sets is just used for measuring the performance of the classifier and is different from the training goal.so for making the best-predicting models it is not used. (Larose, D. T 2014)

So in this paper, the most of data have been used (70 % for training and earning the best predictive models and 30 % of the data to test and measure the performance of the model.)

4.4. Evaluation Methods

Factors used to analyze the proficiency of algorithms include the Accuracy (ACC) metric, and also Area Under Curve (AUC), and Receiver Operating Characteristics (ROC) criteria, which are described below:

Accuracy: Is the ratio of True predictions number to the entire number of input elements. The formula is brought as:

$$\text{Accuracy}= \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where:

TP= True-Positive = number of individuals with diabetes disease who have a positive test outcome
TN= True- Negative = number of individuals without diabetes disease who have a negative test outcome
FP= False-Positive = number of individuals, without diabetes disease who have a positive test outcome
FN= False-Negative = number of individuals, with diabetes disease who have a negative test outcome
ROC= Receiver Operating Characteristics = a plot of True Positive Rate (TPR) vs. False Positive Rate (FPR)
AUC = the Area Under the Curve of ROC

5. Results

In this study, diabetes using several classification algorithms and a combined proposed algorithm called MVO-MLP have been predicted.

The evaluation methods are AUC (Area under the curve) and ACC (Accuracy).

The results were according to the average of ten replications in each algorithm, which are brought in Table 3 and Table 4. Both Table 3 and Table 4 show the improved percentage of MVO--MLP compared to MLP, which the ACC improvement is 107% and AUC improvement is 0.07%.

Table 3. The results according to the average of ten replications in the ACC algorithm

Algorithms	Average of ACC
Naïve Bayes	0.75
LR	0.7743
MLP	0.747
SMO	0.7674
J48	0.7316
RF	0.7651
RT	0.7025
MVO-MLP	0.8541

Table 3 shows the average numbers obtained in each algorithm in the diagnosis of diabetes after ten repetitions in the ACC criterion. As it is clear from the obtained numbers, the approximate average obtained in the ACC criterion in the presented model is 0.8541%, which is an acceptable value. It shows the difference compared to other algorithms. Also, this value has improved by 107% compared to the MLP algorithm itself.

Table 4. The results according to the average of ten replications in the AUC algorithm

Algorithms	Average of AUC
Naïve Bayes	0.808
LR	0.830
MLP	0.793
SMO	0.713
J48	0.730
RF	0.827
RT	0.671
MVO-MLP	0.863

Table 4 shows the average numbers obtained in each algorithm in the diagnosis of diabetes after ten repetitions in the AUC criterion. As it is clear from the obtained numbers, the approximate average obtained in the AUC criterion in the presented model is 0.863%, which is an acceptable value. It shows the difference compared to other algorithms. Also, this value has improved by 0.07% compared to the MLP algorithm itself.

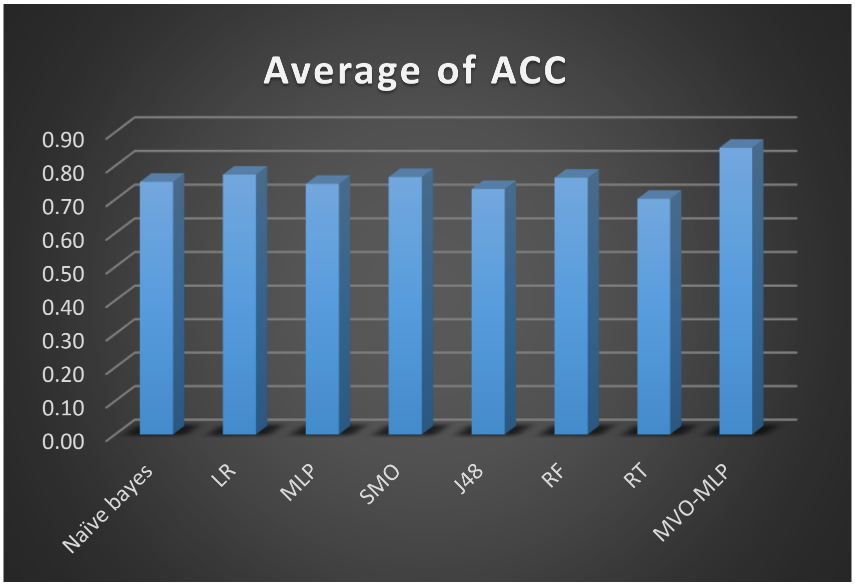


Figure 3. The results of ACC of algorithms in predicting diabetes.

Figure 3 shows a complete view of the results obtained in Table 3. As we can see in this figure, the results obtained after ten repetitions of the algorithms used in the ACC criterion show that this value in the proposed model (MVO-MLP is higher than other algorithms. and, it has increased by 107% Compared to the MLP algorithm.

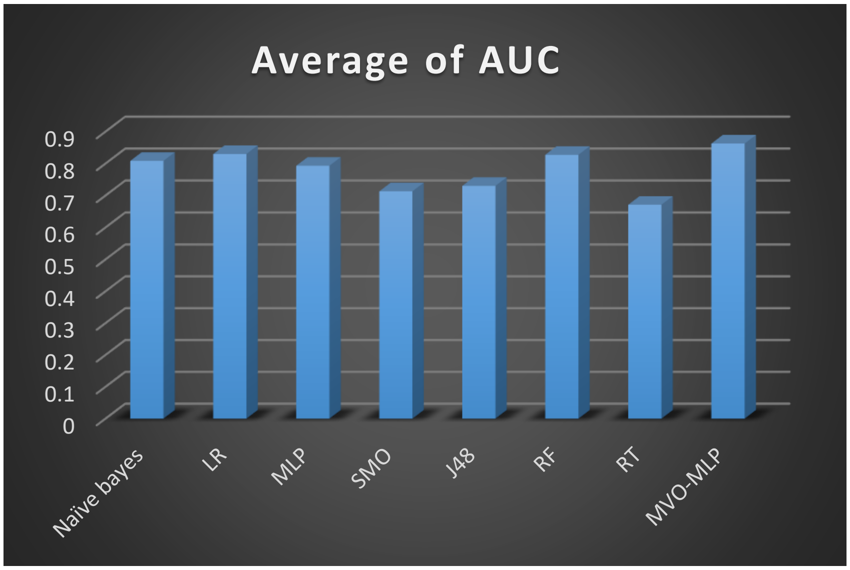


Figure 4. The results of AUC of algorithms in predicting diabetes

Figure 4 shows a complete view of the results obtained in Table 4. As we can see in this figure, the results obtained after ten repetitions of the algorithms used in the AUC criterion show that this value in the proposed model (MVO-MLP is higher than other algorithms. and, it has increased by 0.07% Compared to the MLP algorithm

5.1 ROC Curve

A receiver operating characteristic curve, or ROC curve, is a visual scheme that indicates the execution of a classification pattern at changing limit rates.

The ROC curve is the scheme of the true positive rate (TPR) in return for the false positive rate (FPR) for each of the limit settings. In the interpretation of the ROC diagram, it can be said that the points that are above the median have good performance and the AUC value is above 50%. AUC at points on the median is equal to 50%, and points below the median are in an unfavorable position with an AUC of fewer than 50%. Another significant matter about ROC is the total area below the curve in which the higher one is the better one. Figure 5 and Figure 6 indicate the importance of the location of points and areas below the curve in ROC.

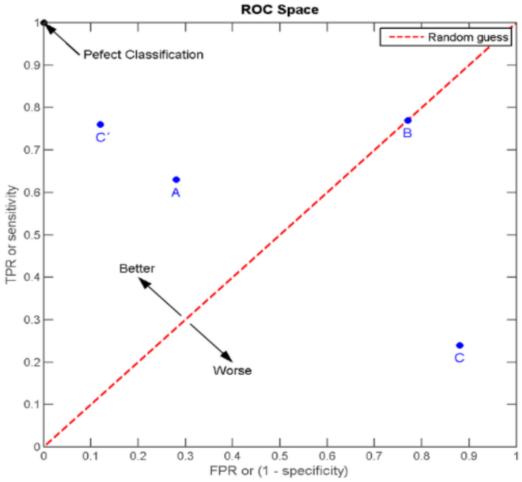


Figure 5. The effects of Distribution of points on ROC

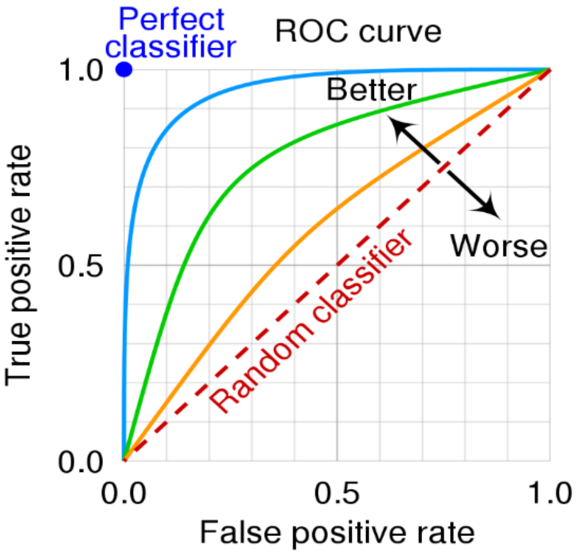


Figure 6. The effect of the area below the curve

Figures 5 and 6 show the effect of distribution points and the area below the roc curve. which are described below:

A ROC area is specified by FPR and TPR as x and y axes, respectively, which indicate pertinent trade-offs between true positive (benefits) and false positive (costs). As TPR equals sensitivity and FPR is tantamount to 1 − specificity, the other name of the graph of ROC is the sensitivity vs (1 − specificity) plot.

The best feasible method of diagnosis would Take place in the superior left angle or coordinate (0,1) of the ROC area, with 100% sensitivity (no false negatives) and 100% specificity (no false positives). Another name of The (0,1) point is a perfect classification. There would be a guess along a diagonal line (the so-called line of no-sensorial) from the bottom left to the top right angle (based on the positive and negative rates). A sensational example of random assuming is a flipping coin decision. With enhancing the size of samples, a random classifier's ROC inclines to the oblique line. When there is a balanced coin, it will incline to the point (0.5, 0.5).

The oblique line separates the ROC area. Points upper than the oblique line indicate good classification output (better than random) and the points under the line indicate not good results.

So It can be concluded from comparing Figure 5 and Figure 6 that the points above the median represent the higher area under the curve and better proficiency. There is also another chart called linear forecasting which predicts the procedure of each curve on the future the other hand if the linear forecasting chart of each curve is ascending it will be evidence of having better efficiency because of the possibility of having a higher area under the curve. Or if it has a descending process the AUC is going to be decreasing in the future

The results of the AUC criterion are given in the form of a ROC diagram of 10 iterations for each algorithm in Figures 7 to 15 and a total ROC of all 8 algorithms in Figure 16, which more smoothly displays the superiority of the suggested algorithm in the AUC criterion.

As can be reflected in Figures 7 to 15 and more clearly from Figure 16, the area below the diagram in the MVO-MLP algorithm is larger than other algorithms, or most parts of its curve are above the median line. So it could be a good document to show the proposed algorithm as the best one on proficiency.it is also obvious from the linear forecasting chart that the process of this curve is going to be ascending with more replications. But other algorithms have subtracting flows.

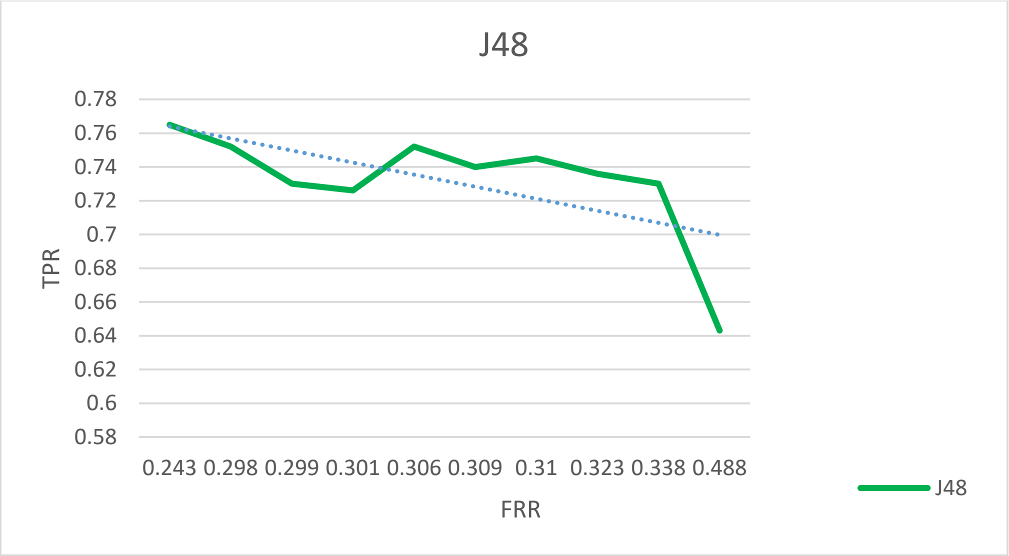


Figure 7. ROC of J48

figure 7 shows the area under the ROC curve of the J48 algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

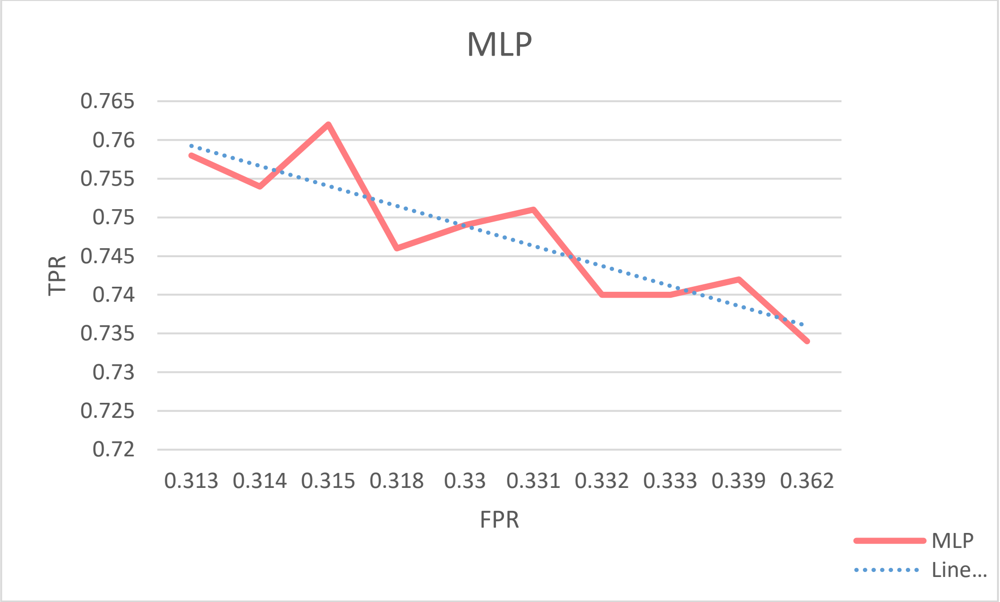


Figure 8. ROC of MLP

Figure 8 shows the area under the ROC curve of the MLP algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

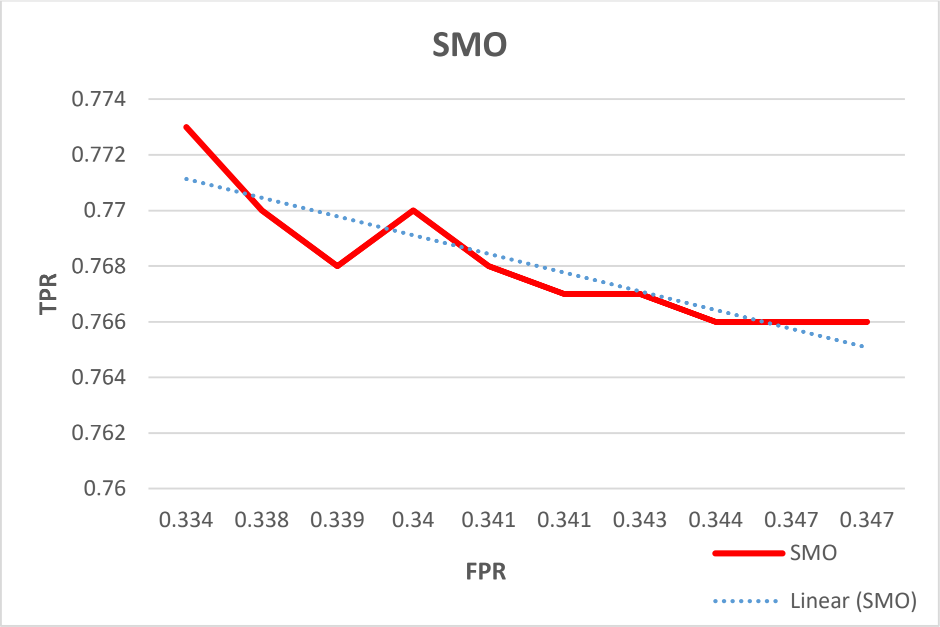


Figure 9. ROC of SMO

Figure 9 shows the area under the ROC curve of the SMO algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

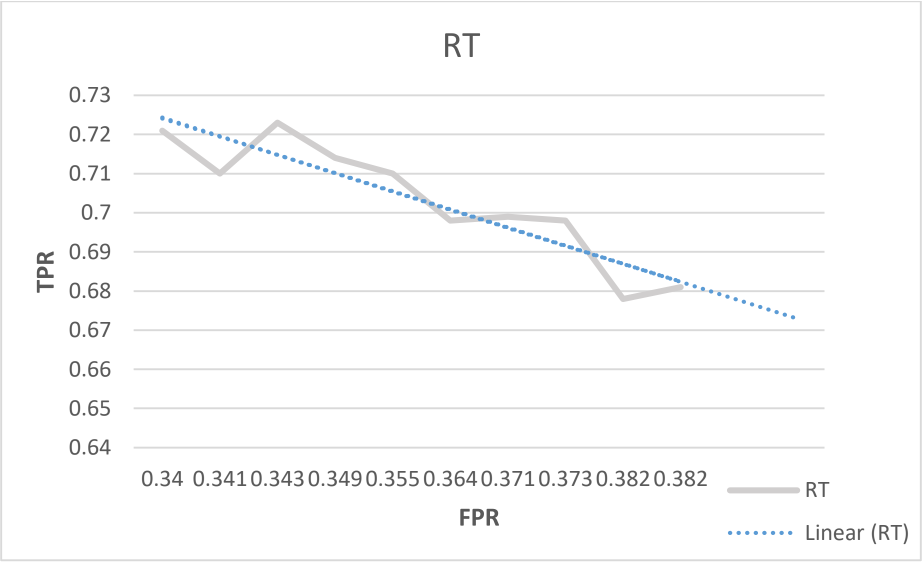


Figure 10. ROC of RT

Figure 10 shows the area under the ROC curve of the RT algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

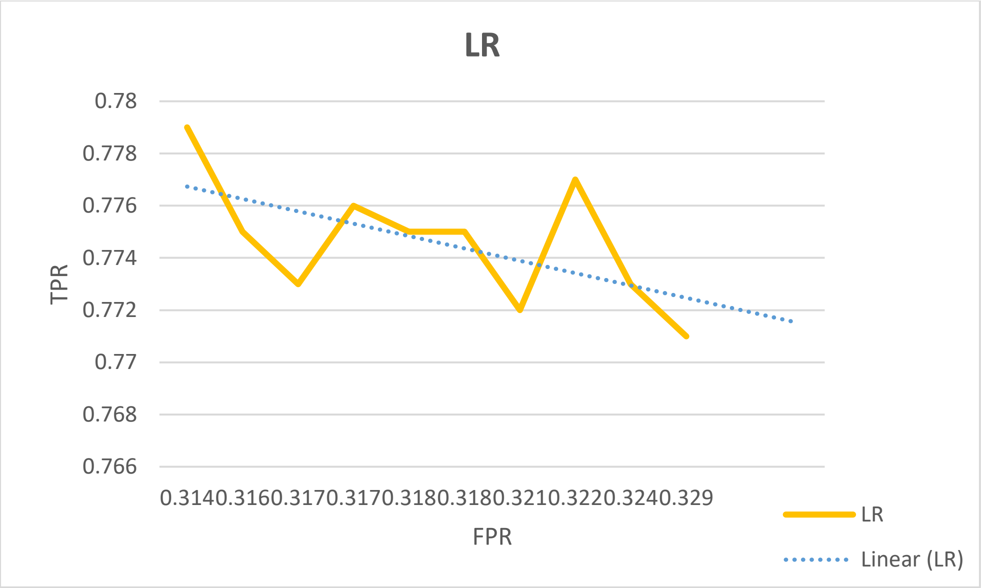


Figure 11. ROC of LR

Figure 11 shows the area under the ROC curve of the LR algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

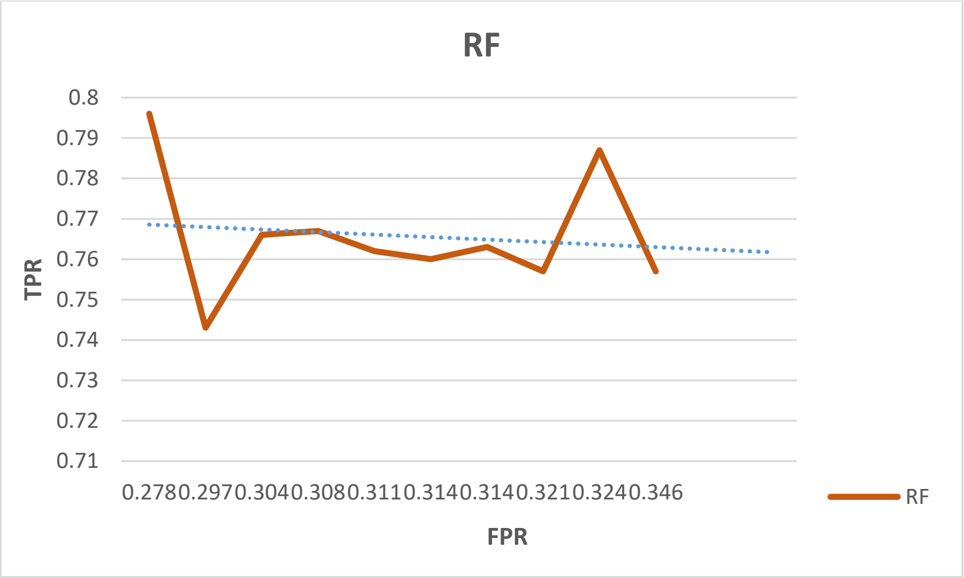


Figure 12. ROC of RF

Figure 12 shows the area under the ROC curve of the RF algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

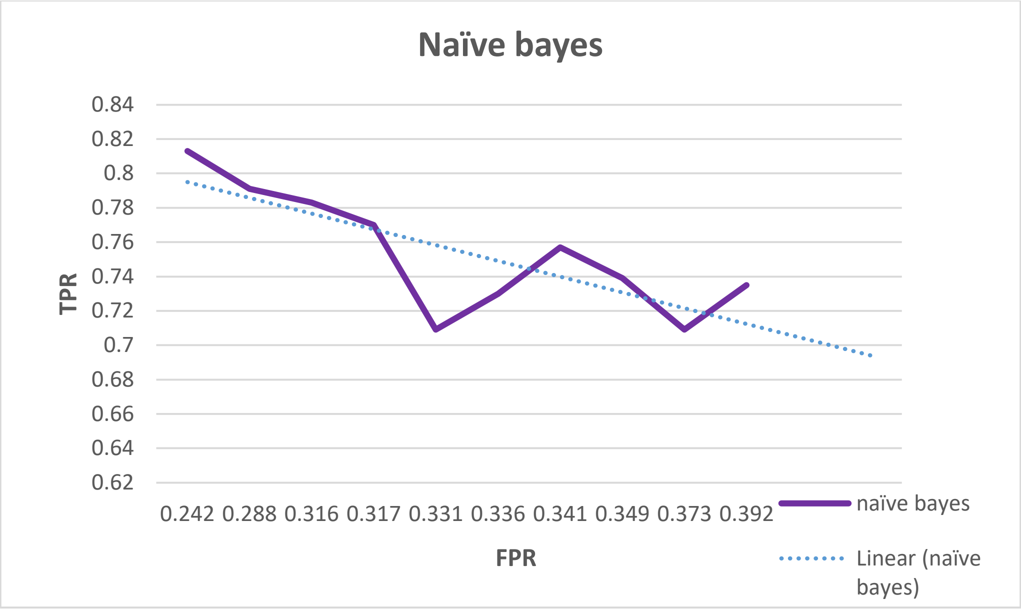


Figure 13. ROC of Naïve Bayes

Figure 13 shows the area under the ROC curve of the Naïve Bayes algorithm in ten repetitions. As it is clear from the figure, the area under the curve is decreasing in more repetitions, which indicates less accuracy in more repetitions.

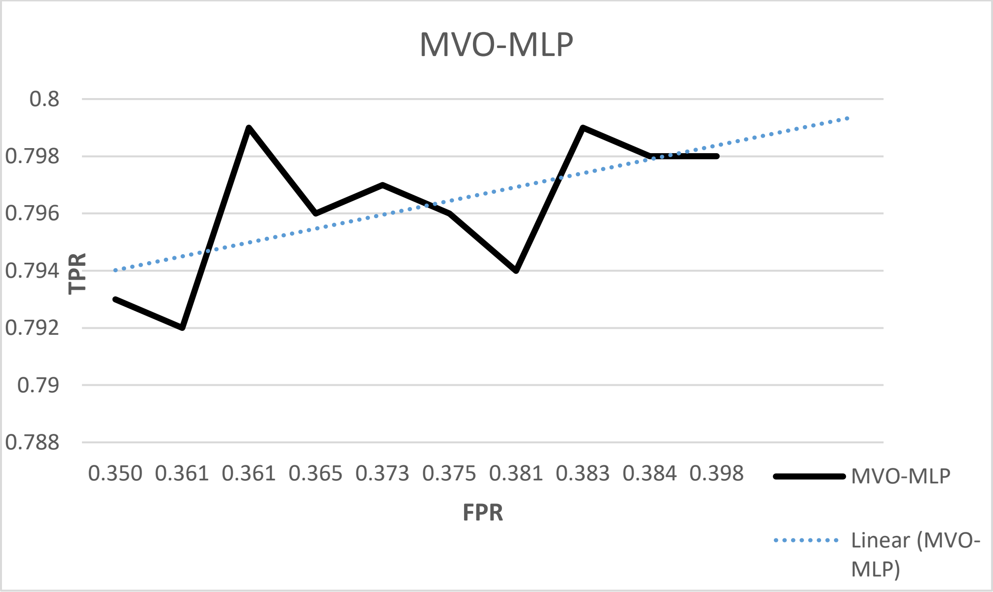


Figure 14. ROC of MVO-MLP

Figure 14 shows the area under the ROC curve of the MVO-MLP algorithm in ten repetitions. As it is clear in the figure, the trend of the chart is ascending, which indicates that the area below the ROC chart and the level off accuracy increases in more repetitions.

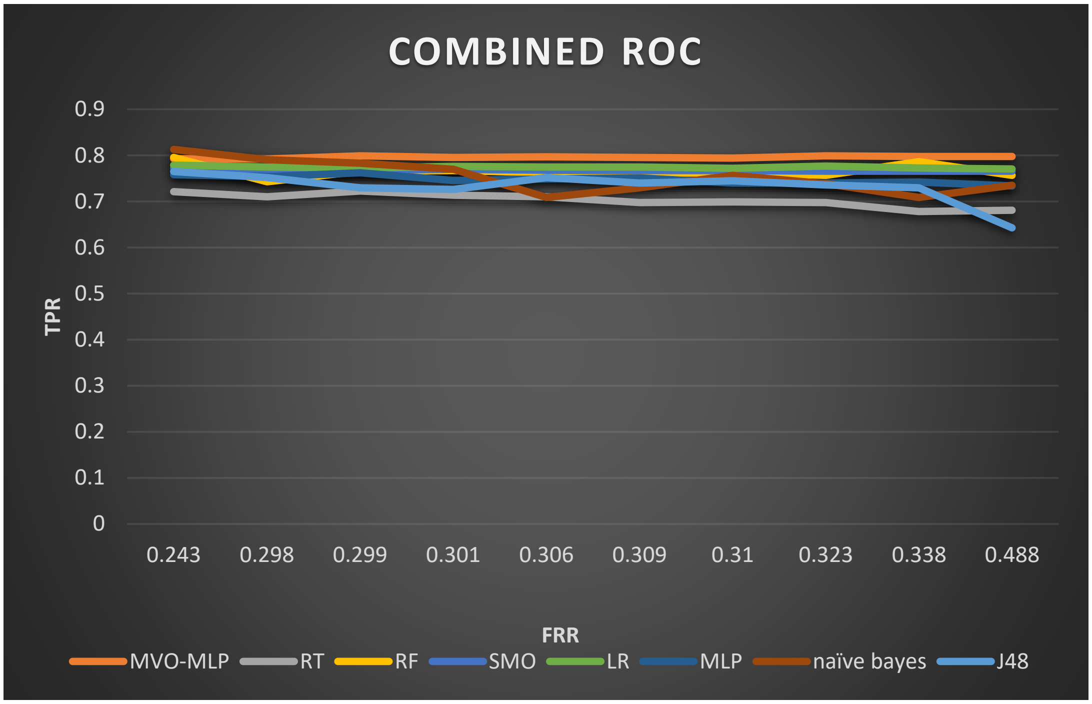


Figure 15. Combined ROC of our algorithms

Figure 16 represents the ROC diagram of all the algorithms which have been used in this paper. The graphical view of the chart can help us compare the performance of the algorithms with the measure of the area under the ROC chart. As said before, the area under the curve for, MVO-MLP is impressive.

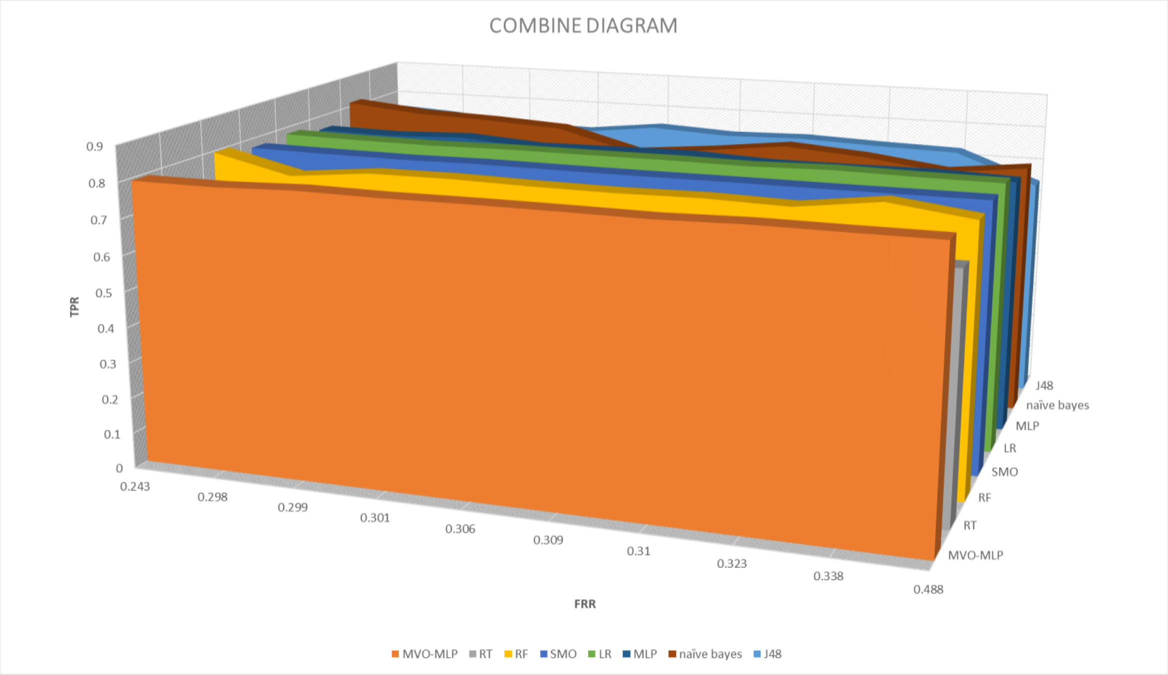


Figure 16. The ROC diagram

6. Conclusion and Future Work

In this paper, several classification machine learning algorithms such as Multi-Layer Perceptron (MLP), Sequential Minimal Optimization (SMO), Naïve Bayes, Logistic Regression (LR), J48, Regression Tree (RT) algorithms, and also Random Forest (RF) have been reviewed in order to predict diabetes by comparing the performance of them.

Moreover, a new hybrid model called MVO-MLP has been proposed which uses the Multi-Verse Optimizer (MVO) algorithm as a tool to train the Multi-Layer Perceptron (MLP) mechanism. In other words, the Backpropagation algorithm is used for training MLP.

In addition, a comparison between the performance of the presented hybrid algorithm with the most proposed applicable machine learning algorithms has been done.

The presented hybrid algorithm in diagnosing and predicting diabetes has been applied using Pima data. Weka and Python have been applied to run the experiments. The number of running each kind of algorithm is 10 and the average of them calculated 30% of the Pima data sets were used for testing and 70% were used to train.

By reviewing the articles and related works done in the previous years, it was discovered that there was no use of new algorithms, and it seems that it is better to use more up-to-date algorithms for better work. Combined algorithms are a good example to test and compare with algorithms. are repetitive and old in predicting diabetes. Therefore, the combined algorithm has been used in this article. By using two important criteria in recent articles, which are ACC and AUC, the disease of diabetes was predicted and by using the new MVO-MLP combination model and comparing it with other used algorithms and the MLP algorithm itself, they found that the prediction accuracy of the combination model The MLP algorithm has improved by 107% in the acc criterion and 0.07% in the AUC criterion compared to the MLP algorithm. Also, these numbers are more than the numbers obtained in other algorithms used in the research. Therefore, the combined algorithm can be introduced as a reliable model for predicting diabetes.

The conducted research can show clear horizons for future works. For example, the effects of using the introduced algorithm can be used in the prediction of other diseases such as cancer. Also, by doing more repetitions in the implementation of the algorithms, the results can be improved. In the future, the use of combined machine learning algorithms will increase until these combined models replace the old classification models.

References

Ahmad, A., Mustapha, A., Zahadi, E. D., Masah, N., & Yahaya, N. Y. (2011). Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. In *Digital Information Processing and Communications: International Conference, ICDIPC 2011, Ostrava, Czech Republic, July 7-9, 2011, Proceedings, Part I* (pp. 537-545). Springer Berlin Heidelberg.Ahmad, F., et al., *Intelligent medical disease diagnosis using improved hybrid genetic algorithm-multilayer perceptron network.* Journal of Medical Systems, 2013. **37**(2): p. 1-8.

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

Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. *IEEE Access*, *10*, 8529-8538.Alam, T.M., et al., *A model for early prediction of diabetes.* Informatics in Medicine Unlocked, 2019. **16**: p. 100204.

Bellamy, L., Casas, J. P., Hingorani, A. D., & Williams, D. (2009). Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The lancet*, *373*(9677), 1773-1779.

Butwall, M., & Kumar, S. (2015). A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. *International Journal of Computer Applications*, *120*(8).

Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, *35*(22), 16157-16173.

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, *2*, 100118.

Cox, M. E., & Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. *Clinical diabetes*, *27*(4), 132-138.

Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D. P. P., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analytics*, *5*, 100301.

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01), 1-16.

Ganie, S. M., & Malik, M. B. (2022). An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators. *Healthcare Analytics*, *2*, 100092.

Goyal, M., Malik, R., Kumar, D., Rathore, S., & Arora, R. (2020). Musculoskeletal abnormality detection in medical imaging using GnCNNr (group normalized convolutional neural networks with regularization). *SN Computer Science*, *1*(6), 1-12.

Gowthami, S., Reddy, R. V. S., & Ahmed, M. R. (2024). Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus. *Measurement: Sensors*, *31*, 100983.

Himsworth, H. P., & Kerr, R. B. (1939). Insulin-sensitive and insulin-insensitive types of diabetes mellitus.

Hina, S., Shaikh, A., & Sattar, S. A. (2017). Analyzing diabetes datasets using data mining. Journal of Basic & Applied Sciences, 13, 466-471.

Islam, M. M., Rahman, M. J., Roy, D. C., & Maniruzzaman, M. (2020). Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(3), 217-219..

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017, February). Predictive analysis of diabetic patient data using machine learning and Hadoop. In 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC) (pp. 619-624). IEEE.

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. Bulletin of Electrical Engineering and Informatics, 12(3), 1728-1737.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

Khaleel, F. A., & Al-Bakry, A. M. (2023). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*, *80*, 3200-3203.

Khan, D. M., & Mohamudally, N. (2011). An integration of K-means and decision tree (ID3) towards a more efficient data mining algorithm. *Journal of Computing*, *3*(12), 76-82.

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). [Retracted] A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of healthcare engineering*, *2022*(1), 1684017.

Kumari, S., Kumar, D., & Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, *2*, 40-46.

Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes research and clinical practice*, *87*(1), 4-14.

Shetty, D., Rit, K., Shaikh, S., & Patil, N. (2017, March). Diabetes disease prediction using data mining. In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)* (pp. 1-5). IEEE.

Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. *Applied Intelligence*, *52*(3), 2411-2422.

Lukmanto, R. B., Nugroho, A., & Akbar, H. (2019). Early detection of diabetes mellitus using feature selection and fuzzy support vector machine. *Procedia Computer Science*, *157*, 46-54.

Marcano-Cedeño, A., Torres, J., & Andina, D. (2011, May). A prediction model to diabetes using artificial metaplasticity. In *International Work-Conference on the Interplay Between Natural and Artificial Computation* (pp. 418-425). Berlin, Heidelberg: Springer Berlin Heidelberg.

El Massari, H., Sabouri, Z., Mhammedi, S., & Gherabi, N. (2022). Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT Standardization*, *10*(2), 319-337.

Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Computing and Applications*, *27*, 495-513.

Nithya, B., & Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 492-499). IEEE.

Olokoba, A. B., Obateru, O. A., & Olokoba, L. B. (2012). Type 2 diabetes mellitus: a review of current trends. *Oman medical journal*, *27*(4), 269.

Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for type-2 diabetic patients. *Expert systems with applications*, *37*(12), 8102-8108.

Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital signal processing*, *17*(4), 702-710.

Rawat, V., Joshi, S., Gupta, S., Singh, D. P., & Singh, N. (2022). Machine learning algorithms for early diagnosis of diabetes mellitus: A comparative study. *Materials Today: Proceedings*, *56*, 502-506.

Samsel, K., Tiwana, A., Ali, S., Sadeghi, A., Guergachi, A., Keshavjee, K., ... & Shakeri, Z. (2024). Predicting depression among canadians at-risk or living with diabetes using machine learning. *medRxiv*, 2024-02.

Theerthagiri, P., Ruby, A. U., & Vidya, J. (2022). Diagnosis and classification of the diabetes using machine learning algorithms. *SN Computer Science*, *4*(1), 72.

Wilson, R. A., & Keil, F. C. (1999). The MIT encyclopedia of the cognitive sciences. A Bradford book..

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, *9*, 515.