Decision Support System for Diabetes Mellitus through Machine Learning Techniques

Tarik A. Rashid
Software & Informatics Engineering
College of Engineering
Salahadin university-Erbil
Hawler, Kurdistan

Saman . M. Abdulla
Software Engineering
College of Engineering
Koya university
Hawler, Kurdistan

Rezhna . M. Abdulla
Software & Informatics Engineering
College of Engineering
Salahadin university-Erbil
Hawler, Kurdistan

***Abstract*—recently, the diseases of diabetes mellitus have grown into extremely feared problems that can have damaging effects on the health condition of their sufferers globally. In this regard, several machine learning models have been used to predict and classify diabetes types. Nevertheless, most of these models attempted to solve two problems; categorizing patients in terms of diabetic types and forecasting blood surge rate of patients. This paper presents an automatic decision support system for diabetes mellitus through machine learning techniques by taking into account the above problems, plus, reflecting the skills of medical specialists who believe that there is a great relationship between patient's symptoms with some chronic diseases and the blood sugar rate. Data sets are collected from Layla Qasim Clinical Center in Kurdistan Region, then, the data is cleaned and proposed using feature selection techniques such as Sequential Forward Selection and the Correlation Coefficient, finally, the refined data is fed into machine learning models for prediction, classification, and description purposes. This system enables physicians and doctors to provide diabetes mellitus (DM) patients good health treatments and recommendations.**

Keywords—Diabetes disease; Blood sugar rate and symptoms; ANN; Prediction and Classification models

I. INTRODUCTION

The International Diabetes Federation stated that within the next 20 years, the figure of diabetic persons will stretch to 285 million in the world [1, 2]. Consequently, numerous research works have been conducted to analyze and categorize the DM patient types [3, 4]. Most of researchers have depended further on artificial intelligence (AI) and data mining (DM) techniques for constructing their classifier or forecaster models. They aimed at targeting two important objectives to AI classifier models; first is to point out the most related features and predictors or statically so called independent variables that should have no correlation among each other and have strong correlation with the desired target. Second is to select a suitable AI technique as a classifier or predictor tool which would possibly produce highest accuracy rate [5, 6]. Thus, at this stage, most of AI models would not provide or improve something to the knowledge of the physicians and medical staffs who are observing DM cases. The only support that they can provide is to categorize the type of DM cases or predict glucose rate in the blood. On the other hand, the most important missing benefit to the physician staffs and even to the indigent patients themselves is to describe the future of DM patients. Thus, it is so crucial to study the symptoms of DM patients not only to categorize their types, but also to envisage what side-effects or more chronic diseases a patient should anticipate.

For the above reasons, the influences of this work can go further than just classifying DM cases. Thus, the main contributions are as follows: 1) It utilizes some independent variables (which are consisted of; a- independent variables that a consultant for DM has considered, b- independent variables that considered by researchers in their previous works and c- independent variables that considered by this work) to diagnose or predict the rate of blood surge for patients through ANN model. 2) After diagnosing or prediction, the work utilizes more variables (symptoms of the patients) and the predicted blood sugar rate for the same patients to find out the relation between the symptoms and five major chronic diseases that diabetic patients have high probability to get them.

The rest of this paper is structured as follows: The next section describes the background of DM and AI techniques. Section 3 describes the proposed method, and in Section 4 the discussion and conclusion of the paper are outlined, and finally, the future work is suggested.

II. DM AND AI TECHNIQUES

The most important AI techniques that have been used by researchers are Artificial Neural Network (ANN), Support Vector Machine, Fuzzy Logic systems, K-mean classifier, and many others [3, 7-11]. ANN is considered to be the most popular one among all. A review work on using ANN in medical diagnoses is accomplished by [7], and it has been displayed in [16] that ANN can have several practices and can have different algorithms for training. Most of research works utilized the multilayer ANN with feed-forwarded back propagation (FFBP) algorithms to achieve DM classification. A research work has been done by [5] to categorize diabetic patients into insulin and non-insulin. The work depended on datasets collected in India, called Pima Indian Diabetes Dataset [8]. Another research work used the same FFBP algorithm to diagnose the DM cases [6]. They collected the database from Sikkim Manipal Institute of Medical Sciences Hospital, Gangtok, Sikkim for the diabetic patients.

Fuzzy logic classifier model is another type of AI tools that has been utilized by researchers [11] to categorize cases into type-1 and type-2. Their work relied on a secondary type of

database called Pima dataset. The accuracy of their work was evaluated against the rate of misclassified cases. A particular work utilized Decision Tree (DT) which is also considered as an AI tool for diagnosing diabetes to achieve classification and compared to ANN. They concluded from their results that DT demonstrated better accuracy [13].

Several approaches and algorithms were used to extract hidden information from biomedical datasets. A research work conducted to classify diabetes cases using Principal Component Analysis and Neuro-Fuzzy Inference. The diabetes disease dataset that used in this study was taken from Machine Learning Database (Department of Information and Computer Science, University of California) and the obtained classification accuracy was 89.47% [14]. Another work conducted to classify diabetes cases and they obtained 78.4% classification accuracy with 10-fold cross-validation (FCV) using Evolving Self-Organizing Map [15]. A combination approach was followed to combine Quantum Particle Swarm Optimization (QPSO), Weighted Least Square (WLS) and Support Vector Machine to diagnose Type-II diabetes. More Research works recorded in this area as the one that applied c4.5 algorithm for classification and it obtained 71.1% accuracy rate [16].

As mentioned in all above works, researchers continuously were busy to classify or diagnose diabetes cases into some predefined categories. Nevertheless, lately, physicians and doctors are concerned about the likelihoods of more chronic diseases or problems that might attack the diabetic patients. Thus, this work is given more descriptive information about diabetic patients through predicting which chronic diseases (problems) more probably can attack a diabetic patient or case based on their detected symptoms.

III. The System Struture

The new proposed approach in this work has a combination form in which the flow operation is shown in Fig. 1. Details on the pro-posed system are elaborated in the next four parts:-

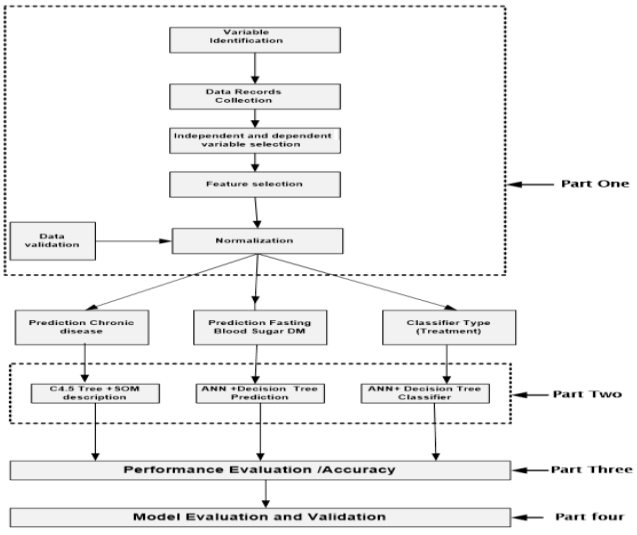


Fig. 1. The Flowchart of the Proposed Emotion Recognition System Structure

A. Part one: Data Collection and Pre-Processing

The part one can be divided into the following stages:-

1) Model's Variables Identification

From the relevant works that mentioned in section 2, many independent variables that are related to DM cases could be obtained. However, it is important to mention that each work has utilized the variables in a specific way. The collected independent variables can be grouped into physical, biological, and symptoms. Most physical records were about the height, weight, mass index, age and sex of the DM patients [17]. There are also some physical recodes that usually considered for specific cases, for instance, the number of pregnancy for a DM patient was only considered for female cases or if the research study is about Gestational Diabetes Mellitus GDM [13]. The biological related records are mostly Blood Glucose rate, Systolic blood pressure, and Diastolic blood pressure [6]. Researchers utilized these independent variables to build models that can classify DM cases or predict rates of glucose in blood.

The final group of independent variables is known as patient's symptoms. These records are more related to troubles and problems that a DM patient may suffer form, and together some of them are related to signs of DM disease. Some of these problems or symptoms are related to some chronic diseases. The second source of this work is to check the record list of DM patients in a specialist clinic. To achieve that, an official communication conducted between Salahaddin University-Erbil, College of Engineering and Layla Qasim Consultant Center Clinic for DM through the Department of Health in Erbil. Table 1, shows the most important records (independent variables) that a specialist clinic or consultant for DM has considered in his documentations. The table is also showing the most important independent variables that considered by researchers in their works. The third column in the table is about the independent variables that considered by this work to build an enhanced model for diagnosing and describing DM patients. The column shows how this work included the most important variables that are considered by DM consultant clinics and previous researches in DM diagnoses filed.

2) Data and Records' Collection

The data collection in this research work involves 26 variables through the process of building and simulating an enhanced medical model for diagnosing and describing DM patients. At this stage, it is necessary to collect records and data about all those variables. To achieve that, a second visit to Layla Qasim Center Clinic for DM patients has been made.

Through different processes, tests and interviewing, the record fields for each (501 patients) patient have been collected. The process of collecting data and records has been achieved carefully by the authors of this work under the supervision of specialist, medical and physician personnel.

After 60 days of getting data and records about DM patients. Two types of records obtained; the first type is those variables that have units such as Weight (kg), Height (cm), and S.BP (mg/dl). The second type is those variables that are logically recorded as 0 or 1. These logically variables are

related to some symptoms that DM patients might have felt them, such as Polyuria, Weakness, Numbness, and Coma.

3) Dependent and Independent Variable Selection

All records (there are 26 variables) that have been collected for DM patients are tabulated for each patient. It is necessary now to define the independent variables and dependent variables among these records. It is also necessary to find out the correlation between the dependent and independent variables in order to define the targets for both prediction and description sub-models.

4) Feature Selection

This work utilized two types of feature selection methods. Details of feature selection techniques and their implementation are outlined in the three points below:-

- Sequential Forward Selection (SFS): It attempts to find the best feature subset that decreases the feature space dimensionality with the smallest loss in classification accuracy. In other words, for a set of *D* features, the algorithm chooses a subset of size *d* <*D*, which has the greatest ability to discriminate between classes. The goodness of a particular feature subset is evaluated using an objective function, $J(Y_m)$, where Y_m is a feature subset of size *m* [18]. SFS is considered as a greedy search algorithm that chooses a top set of features for extraction through beginning from a void set and successively adding a distinct feature in the superset to the subset when increasing the value of the chosen objective function [19]. This type of algorithm has $O(n^2)$ worst-case complexity. Suppose we have a set of d_1 features X_{dI}, for each of the feature yet not selected ξ_j (i.e. in X- X_{dI}) the criterion function is evaluated according to the below equation [19]:-

$$J_j = J(X_{dI} + \xi_j) \quad (1)$$

The feature that yields the maximum value of J_j is chosen as the one that is added to the set X_{dI}. Thus, at each stage the variable is chosen, when added to the current set, and it maximizes the selection criterion. The feature set is initialized to the null set. Whenever the best improvement makes the feature set worst, or when the maximum allowable number of feature is reached, the algorithm terminates. Here, *J* can be given by the below equation [20]:-

$$j = X_k^T . S_k^{-1} \quad (2)$$

Where X_k is a *k* dimensional vector and S_k is a *KxK* positive definite matrix, where *K* features are used. At each stage of the search; sets of subsets are generated for evaluation. Variable ξ_j is chosen for which $J(X- \xi_j)$ is the largest. The new set is $(X- \xi_j)$. This process is repeated until the set of required cardinality remains. The following algorithm explains the whole procedure:-

a) initial X_{dI} *= null, m=1,*

X: is defined as a full set of feature

b) Choose new feature (ξ_j)

$$\xi_j = X - X_{dI} \quad (3)$$

c) K-fold checking learning performance

$$Avarage\ J_m = X^T . X_{dI} . \xi_j \quad (4)$$

d) Checking feature

$Avarage\ J_{m+1}$ *is better than* $Avarage\ J_m$

e) Go to step 2

Worst set of X_{dI} *is achieved*
Number m features are reached

f) Create training and testing sets

- The Correlation Coefficient *(r): r*, is a rapid portion that can define the scope of the statistical correlation between two variables. *r,* is scaled in a way that is constantly between -1 and +1. As soon as *r* is close to 0, it means that there is little correlation between the variables and the farther away from 0, *r* is, in either the positive or negative direction, the greater the correlation between the two variables. To compute the correlation between two variables, below steps can be followed [23]:-

a) Start with a set of data, x and y points. Each data point is kept in a separate row.

b) Find $\bar{x}$, $\bar{y}$, *the mean of x and y respectively. To do this, add the values of x and divide by the number of points; then, do the same process for y. Use equation below: - .*

c) $\bar{x} = \sum x/n$, $\bar{y} = \sum y/n$ (5)

d) Subtract $\bar{x}$ *from each value of x and subtract* $\bar{y}$ *from each value of y to get a new table of rows.*

e) Where, $y - \bar{y}$, $x - \bar{x}$

f) Compute the products of each row in step 3 and calculate the sum.

g) Take each x value in step 3, square it and calculate the sum of all points; do the same thing for y.

h) Calculate the square root of the product of the sums of the squares in step 5.

i) Calculate r by dividing the sum in step 4 by the value in step 6.

$$r = \frac{\sum (X - \bar{X}).(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 . (Y - \bar{Y})^2}} \quad (6)$$

The most significant variable of these is added to the model, so long as its *P-value* is below some pre-set level. It is customary to set this value above the conventional 0.05 level at say 0.10 or 0.15, because of the exploratory nature of this method. The number of variables that fed into feature selection technique are 24 variables, assume it as *X*. The first step is to check whether the validation of data is acceptable or not. This work utilized holdout validation method as a function, which depends on checking the re-substitution performance rate. The validation has been checked using Quadratic Discrimination Analysis *(QDA)* to check the covariance matrix of training set whether they are positive or negative. The acceptable value of the *(QDA)* should be positive. Assume V_{QAD} is the function of the validation check of features *X*.

$$P\ value = V_{QAD}(X) = Positive\ number \quad (7)$$

The first check that is done by this work is the value of the function of V_{QAD}, which represented as *P* value for variables.

The second check is doing t-test, which defines the relation between each *P* value that found for each variable and Cumulative Distribution, which denoted as *CDF* (cumulative distribution function), for each *P*. The relation between these two values should be as close as to one. For any variable, if the relation value is close to zero, the variable will be considered as insignificant. Or else, the variable will be significant if the *CDF* value of *P* is close to one.

Features are selected based on the correlation coefficient value between independent variables and target variable. Although, the correlation coefficient usually used to find the relation between features themselves, in this work it has been utilized to find the correlation between each features and the target variables. The feature selection implementation can be explained as follows:-

- Feature Selection Implementation for DM classification Sub model: The input data set includes 24 variables, excluding the Diabetes Type and the FSR columns as they are considered as targets or dependent variables. Based on the SFS algorithm, the significant features will be those variables that their P-Values are less than 0.09. Table 2, shows the output of the SFS and the features that can be selected for DM classification, which used to distinguish Type-I DM from Type-II. According to the *P* values of the all 24 variables, the won variables are the 16 variables that are highlighted in yellow color in the table. For the second test, this work has found the relation between *P* values and *CDF (P)* values. As explained, the value of this function should be as close as possible to one. According to Fig. 2, which shows the *P-CDF(P)* relation of features that collected by this work, is 33 % of all features, which is equal to eight features that are insignificant. However, the remaining 66 % which is equal to 16 features are significant.
- Feature Selection for Blood Surge Prediction Sub model: The P-Value in this case has the same range (only less than 0.09 is considered). Table 3 illustrates that P-value of all variables and the highlighted values are the considered features. According to the condition, only 14 features considered as significant variables.

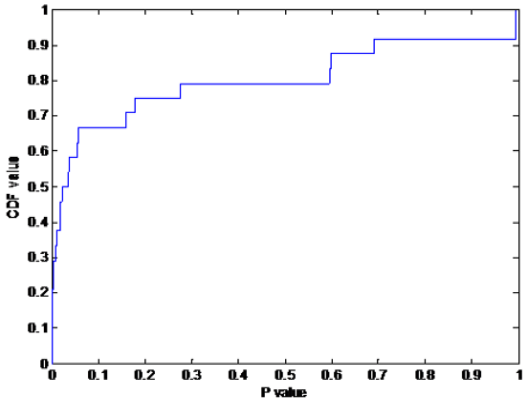


Fig. 2. P-CDF (P) relation of features with P value

5) Data Normalization

Techniques such as Min-Max and Z-Score are used to normalize data. The Min-Max involves the linear transformation on raw data. *Min A* and *Max A* are minimum and maximum value for the attribute *A* respectively. This technique maps the value of attribute *A* into range of [0,1], as in the following equation [21, 22]:-

$$v' = \frac{v - MinA}{MaxA - MinA} \qquad (8)$$

Z-Score is considered as a useful normalization technique when the actual minimum and maximum values of attribute are unknown. This is expressed in the below equation [22]:-

$$v' = \frac{v - \bar{A}}{\sigma A} \qquad (9)$$

Where, $\bar{A}$, σA and v are the mean, standard deviation and value of attribute of *A* respectively. The main advantage of this technique is to put the row data in specific range, so that models can map input-output relationship easily. The main problem that should be avoided in such process, is occurring zero redundant inside a specific attribute.

6) Data Validation

After normalizing input data, it is necessary to validate the data, to observe whether they are generalized or not. This work executed the 10-fold method to discover the performance of each fold of data. Both obtained normalized tables have been fed to the proposed ANN with different division of training and testing data sets. For both tables, same steps are followed. The process started to get 10 % of all data, which is 50 observations, as a testing part of data and the remaining 451 observations are used for training. Each time, the selected data for testing have been changed to another 50 observations. Through ten times, all data are used for testing and training. Tables 4 and 5 are illustrating the results of this 10-fold validation process. All performance records showed that collected data are validated data as there is no outlier performance among the records.

B. Part two: DM Medical Model

The main model has three sub models. The first part receives five independent variables for each patient, and it does the prediction and / or classification. The second sub-part is more important than the first one as it provides more important information to physician staffs. Details of each sub models are as follows:-

1) The classification sub model

An ANN is proposed to build a classifier model that can distinguish the type of treatment between insulin and noninsulin. Matlab program is used to build the DM treatment based ANN classifier sub model. The network has been trained based on back propagation algorithm, the network receives the 11 features [3, 24].

2) FBS prediction for DM patients

ANN is not only used for classification, it can be adopted for prediction too. In this work, the same algorithm is used for training the ANN to forecast the FBS rate for DM patients. The

only thing that changed with the presented ANN is the number of the selected significant features [3], the network receives 10 features.

3) DM description sub model

The last sub part of the main medical model is giving more details and information on a DM case with reference to symptoms with common chronic diseases relationships. The input part of this sub model is involved 10 features (Symptoms) and the output is 6 chronic diseases and problems, which are more common among DM cases. Through this prediction part, it will be easy for any physician to provide more details and information about the problems and chronic diseases that a DM patient might get them based on his/her symptoms. Two AI models were used for evaluations, these are namely; c45 and Self-Organizing Map (SOM) [7].

C. Part three: Performance Evaluations

There are five measuring parameters in this work, namely; True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), and the accuracy rate. The classification results are shown in Table 6. It can be seen from the table, trial 4 has the best TP and FN records, while trail 6 recorded the best TN and FN. And the best accuracy recorded at the trial 10. This is because, accuracy is directly changed with the ratio of all corrected classified objects to all object's number. The accuracy of the tree prediction sub model is shown in the Fig. 3. The figure shows the relation between a specific cost that required visiting a node and the number of nodes that have the same cost. Because the proposed decision tree is working as a regression mode, thus, this cost is the average squared error over the observations in that node. The dashed line in the cost is representing the minimum cost among the set of costs.

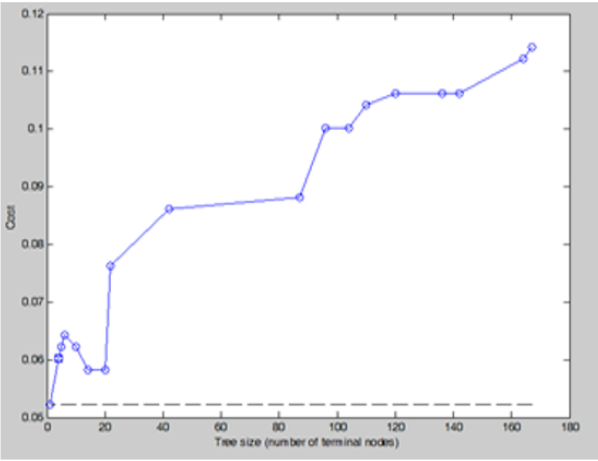


Fig. 3. The cost accuracy of the proposed decision tree sub model

D. Part four: Models Evaluations and Validations

In this work, the proposed model evaluated against supervised and unsupervised AI models. Fig. 4 shows the performance evaluation for classification AI models with / without feature selection process.

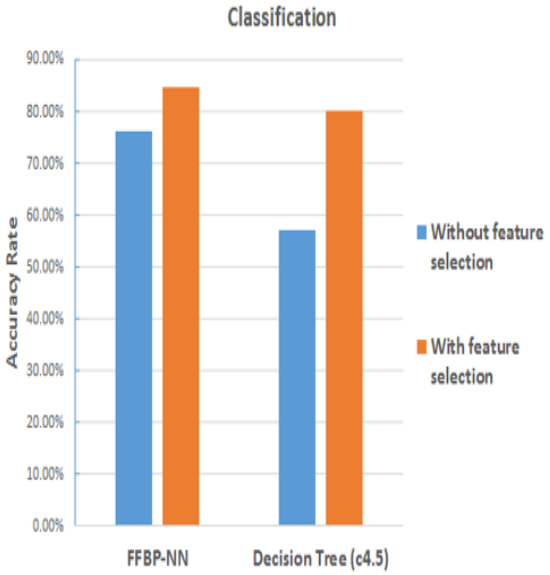


Fig. 4. Evaluation For classification

Fig. 5 show the performance evaluation for prediction FBS with /without feature selection process.

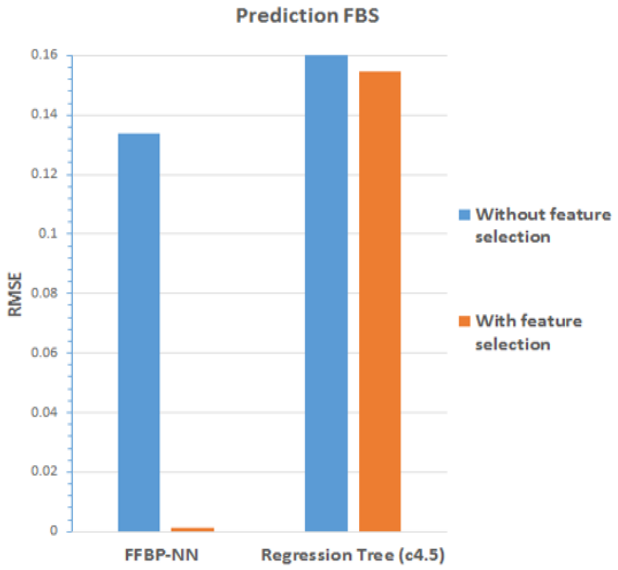


Fig. 5. Evaluation for Prediction

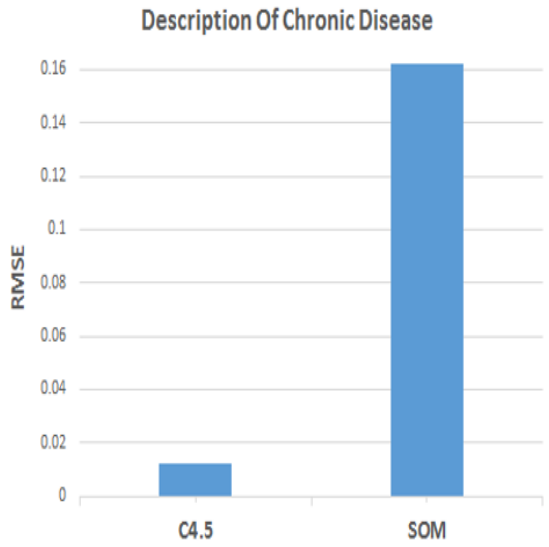
Fig. 6 shows the performance evaluation two AI models for Description of Chronic Disease.

Fig. 6. Evaluation for description

It is evident that ANNs trained with backpropagation learning algorithm using feature selection for the classification type of DM and for prediction FBS are better than c4.5 tree, however, for description of chronic disease the c4.5 is better than SOM.

IV. Discussion and Conclusion

In this paper, three intelligent models are designed for classification, prediction, and description purposes to offer complete knowledge about Diabetes Mellitus patients. Classification and prediction (classifying the types of MD patients and predicting FBS) models are very imperative for DM patients as doctors will be influenced by the outputs of these two models so that to espouse the type and dosage of treatments. Based on some symptoms extra health care recommendations are also specified by physicians to put DM patients away for some potential side effects.



It is worth to say that standard neural networks are good techniques for classification and easier to train when compared with deep neural networks that are regularly more difficult to train. This is a bad news, nevertheless, it is proven that if a deep neural network was, it would be much more controlling and prevailing than a standard neural network [25].

V. Recommendations for Future Work

While the suggested approach using artificial neural networks in this paper delivered encouraging outcomes for classification, prediction, and description purposes. As a consequence, the following future work can be suggested:-

1) It would be a good practice to use Deep Learning algorithms for classification problems instead of standard neural networks.

2) On the other hand, this research work can be upgraded in terms of learning algorithms such as using Grey Wolf Optimizer (GWO) and Bat algorithm (BA) which are freshly suggested swarm-based meta-heuristic.

TABLE I. The type of Variables used in Kurdistan's Clinics, Previous Research Works and in this Research Work

Variables of DM's Documented by consultants (Layla Qasim Clinic)	**Variables in DM Diagnosing and Classification in Previous Research Works**	**Variables in the Proposed Enhanced Medical Model**
Personal	***Personal***	***Personal***
1. Privacy records	1. Age 2. Weight 3. Height 4. Body index 5. Sex 6. Privacy records.	1. Age 2. Weight 3. Height 4. Sex 5. Privacy records.
Diabetic Related Records	***Diabetic Related Records***	***Diabetic Related Records***
1. FBS test 2. RBS test 3. HBA1C test 4. Insulin or tablet base (Treatment)	1. RBS test. (Glucose level) 2. Insulin or tablet (Treatment) 3. Diabetes pedigree function (pedi) (*Inheritance issue*) 4. HBA1C test	1. Fasting BS test. (Glucose level) 2. Insulin or tablet (Treatment) (class A or class B) 3. Diabetes pedigree function (pedi) (*Inheritance issue*) 4. Since when (year base)
Symptoms and problems Related Records	***Symptoms and problems Related Records***	***Symptoms and problems Related Records***
Not found	1. Polyuria 2. Nocturia 3. Polydpsia 4. Weakness 5. Paraesthesia 6. Frequency 7. Weight loss 8. Numbness 9. Polyhagia 10. Coma 11. Thirst 12. VD 13. Imp	1- S. Blood Pressure. 2- D. Blood Pressure. 3- Polyuria 4- Nocturia 5- Polydpsia 6- Weakness 7- Paraesthesia 8- Urinal Frequency 9- Weight loss 10- Numbness 11- Polyhagia 12- Coma 13- Heart Problem 14- Teeth Problem 15- Kidney Problem 16- Eyes Problem 17- Diabetic Foot (injury or damages)

TABLE II. THE P-VLAUE OF FEATURES (VARIABLES) FOR CLASSIFICATION SUB-MODEL

Feature Name	P-Value	Included or Excluded
Sex	0.623	No
Age	1.66 x E-33	Yes
Weight	1.38 x E-08	Yes
Height	0.021	Yes
S.BP	1.36 x E-09	Yes
Inheritance	0.020	Yes
D.B.P	0.021	Yes
Polyuria	0.110	No
Nocturia	0.032	Yes
Polydpsia	0.293	No
thirsty	0.127	No
Weakness	0.082	Yes
Par aesthesia	0.1812	No
Urinal Frequency	0.004	Yes
Losing Weight	0.025	Yes
Numbness	0.002	Yes
Polyhagia	0.915	No
Coma	0.060	Yes
Since When	0.0009	Yes
Eyes Problem	0.0004	Yes
Heart Problem	0.0885	Yes
Teeth Problem	0.0062	Yes
Kidneys Problem	0.9357	No
Injury Probe	0.3991	No

TABLE III. THE CORRELATION-VLAUE OF FEATURES (VARIABLES) FOR PREDICTION SUB-MODLE

Feature Name	P-Value	Included or Excluded
Sex	0.1956	No
Age	0.8169	No
Weight	0.1185	No
Height	0.0346	Yes
S.BP	0.8745	No
Inheritance	0.4134	No
D.B.P	0.0230	Yes
Polyuria	1.18 x E-05	Yes
Nocturia	0.00251	Yes
Polydpsia	1.19 x E-05	Yes
thirsty	5.36 x E-06	Yes
Weakness	0.00860	Yes
Paraesthesia	0.0012	Yes
Urinal Frequency	0.0014	Yes
Losing Weight	0.0615	Yes
Numbness	0.0005	Yes
Polyhagia	0.8836	No
Coma	0.2300	No
Since When	0.3106	No
Eyes Problem	0.0001	Yes
Heart Problem	0.0557	No
Teeth Problem	0.0422	Yes
Kidneys Problem	0.0923	No
Injury Problem	0.0059	Yes

TABLE IV. PERFORMANCE OF PREDICTION FASTING BLOOD SUGAR

Testing Data	Performance
1:50	0.0178
51:100	0.0171
101:150	0.0151
151:200	0.0210
201:250	0.0239
251:300	0.0149
301:350	0.0165
351:400	0.0274
401:450	0.0207
451:501	0.0158

TABLE V. PERFORMANCE FOR CLASSIFICATION TYPE OF DM

Testing Data	Performance
1:50	0.0298
51:100	0.0441
101:150	0.0279
151:200	0.0236
201:250	0.0400
251:300	0.0210
301:350	0.0195
351:400	0.0371
401:450	0.0278
451:501	0.0201

TABLE VI. THE CONFUSION MATRIX FOR THE MD CLASSIFIER SUB-MODEL

#	TP%	TN%	FP%	FN%	Accuracy %
1	148	234	44	75	76.2
	29.	46.7	8.8	15.0	
2	144	242	36	79	77.0
	28.8	48.3	7.2	15.8	
3	124	226	52	99	69.9
	24.8	45.1	10.4	19.8	
4	179	238	40	44	83.2
	35.7	47.5	8.0	8.8	
5	172	245	33	51	83.2
	34.3	48.9	6.6	10.2	
6	133	251	27	90	76.6
	26.5	50.1	5.4	18.0	
7	176	250	28	47	85.0
	35.1	49.9	5.60	9.4	
8	166	245	33	57	82.0
	33.1	48.9	6.6	11.4	
9	172	239	39	51	82.0
	34.3	47.7	7.8	10.2	
10	186	250	28	37	87.0
	37.1	49.9	5.6	7.4	

REFERENCES

[1] P. Srimani and S. Koti, "Medical diagnosis using ensemble classifiers-a novel machine-learning approach," J Adv Comput, vol. 1, pp. 9-27, 2013

[2] S. Wild, G. Roglic, Green, A., et al., "Global prevalence of diabetes estimates for the year 2000 and projections for 2030," American Diabetes Association, Diabetes care, vol. 27, no. 5, pp. 1047-1053, 2004.

[3] A. Kumari and R. Chitra., "Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications, 2013. vol. 3, no. 2, pp. 1797-1801, 2013.

[4] B. Deekshatulu, and P. Chandra., "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection, " Global Journal of Computer Science and Technology, vol. 13, no. 3, 2013.

[5] Z. Zainuddin, O. Pauline and C Ardil, "A neural network approach in predicting the blood glucose level for diabetic patients," International Journal of Computational Intelligence,. vol 5, no. 1: pp. 72-79, 2009.

[6] B. Adeyemo, and E. Akinwonmi, "On the Diagnosis of Diabetes Mellitus Using Artificial Neural Network Models Artificial Neural Network Models," African Journal of Computing & ICT Reference Format, vol. 4, no. 1, pp. 1-8, 2011.

[7] F. Amato, F. Amato, A. López., P-M. María, et al., "Artificial neural networks in medical diagnosis," Journal of Applied Biomedicine, vol.11, no. 2, pp. 47-58, 2013.

[8] G. Karegowda, V. Punya, A. Jayaram et al., "Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4. 5," International Journal of Computer Applications, vol. 45, 2012.

[9] L. Liberti, C. Lavor, N. Maculan, et al., "Euclidean distance geometry and applications," SIAM Review, vol. 56, no. 1, pp. 3-69, 2014.

[10] R. Dey, V. Bajpai, G. Gandhi, et al, "Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus," in IEEE Region 10 and the Third international Conference on Industrial and Information Systems," ICIIS 2008. Kharagpur: IEEE, 2008.

[11] F. Baldwin and W. Xie, "Simple fuzzy logic rules based on fuzzy decision tree for classification and prediction problem," in Intelligent information processing II, Springer. pp. 175-184, 2005.

[12] B. Yegnanarayana, B., "Artificial neural networks, " PHI Learning Pvt. Ltd, 2009.

[13] E. Caballero-Ruiz, E., G. García-Sáez, M. Rigla, et al, "Automatic Blood Glucose Classification for Gestational Diabetes with Feature Selection: Decision Trees vs. Neural Networks," in XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, Springer, 2014.

[14] B. Yegnanarayana1, "Artificial neural networks for pattern recognition Sadhana, vol. 19, no. 2, pp. 189-238, 1994.

[15] A. Feizollah, N. Badrul Anuar, R. Salleh, et al., "A review on feature selection in mobile malware detection. Digital Investigation, vol. 13,pp. 22-37, 2015.

[16] E. Berglund and J. Sitte, J., "The parameterless self-organizing map algorithm. Neural Networks," IEEE Transactions on,. vol. 17, no. 2, pp. 305-316, 2006.

[17] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks," in Data Storage and Data Engineering (DSDE), 2010 International Conference, 2010.

[18] S. García, J. Luengo and F. Herrera, "Feature Selection, in Data Preprocessing in Data Mining, Springer. pp. 163-193, 2015.

[19] M. Fauvel, C. Dechesne, A. Zullo, et al., "Fast forward feature selection for the nonlinear classification of hyperspectral images," arXiv preprint arXiv:1501.00857, 2015.

[20] L. Burrell, G. Georgoulas, E. Marsh, E., et al., "Evaluation of Feature Selection Techniques for Analysis of Functional MRI and EEG," in International Conference on Data Mining 2007, Las Vegas, 2007.

[21] N. Mohd, H. Atomia and Z. Rehman, Z., "The effect of data pre-processing on optimized training of artificial neural networks", Procedia Technology, 4th International Conference on Electrical Engineering and Informatics, ICEEI 2013, vol. 11, pp. 32–39, 2013.

[22] A. Semary, A. Tharwat, E. Elhariri, E., et al., "Fruit-Based Tomato Grading System Using Features Fusion and Support Vector Machine," in Intelligent Systems' 2014, Springer. pp. 401-410, 2015.

[23] R. Momberum, "A comparative Study of the capital structures of liquid and liquidity - stressed Banks," in Financial Managmnet 2012, University of Johannesburg: http://hdl.handle.net/10210/8563 p. 115, 2012.

[24] T. Rashid, S. Abdullah and R. Abdullah, "An Intelligent Approach for Diabetes Classification," Prediction and Description, Editors: Vaclav Sansel, Ajith Abraham, Pavel Kromer, Millie Pant, Azah Kamilah Muda, In book: Series : Advanced in Intelligent Systems and Computing, Edition: 424, Chapter: IBIA 2105 Proceeding, Publisher: Springier Verlag, , pp.323-335, 2015.

[25] N. George, "Deep Neural Network Toolkit & Event Spotting in Video us-ing DNN features," master thesis, department of computer science and engineering, Indian institute of technology madras, 2015.