I.J. Intelligent Systems and Applications, **2024, 1, 11-23**
Published Online on February 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2024.01.02

Detection of Diabetes using Combined ML Algorithm

Shifat Jahan Setu
Department of Computer Science & Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh
E-mail: sjahans.cseju@gmail.com
ORCID iD: https://orcid.org/0009-0009-0454-2521

Fahima Tabassum*
Institute of Information Technology, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh
Email: fahima@juniv.edu
ORCID iD: https://orcid.org/0000-0002-0183-3907
*Corresponding author

Sarwar Jahan
Department of Computer Science and Engineering at East West University, Dhaka, Bangladesh
E-mail: sjahan@ewubd.edu
ORCID iD: https://orcid.org/0000-0002-3080-8145

Md. Imdadul Islam
Department of Computer Science & Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh
E-mail: imdad@juniv.edu
ORCID iD: https://orcid.org/0000-0003-2045-6382

Received: 20 July 2023; Revised: 17 September 2023; Accepted: 07 October 2023; Published: 08 February 2024

Abstract: Recently data clustering algorithm under machine learning are used in 'real-life data' to segregate them based on the outcome of a phenomenon. In this paper, diabetes is detected from pathological data of 768 patients using four clustering algorithms: Fuzzy C-Means (FCM), K-means clustering, Fuzzy Inference system (FIS) and Support Vector Machine (SVM). Our main objective is to make binary classification on the data table in a sense that presence or absence of diabetes of a patient. We combined the four machine learning algorithms based on entropy-based probability to enhance accuracy of detection. Before applying combining scheme, we reduce the size of variables applying multiple linear regression (MLR) on the table then logistic regression is again applied on the resultant data to keep the outlier within a narrow range. Finally, entropy based combining scheme with some modification is applied on the four ML algorithms and we got the accuracy of detection about 94% from the combining technique.

Index Terms: FIS, SVM, FCM, Logistic Regression and Combining Algorithm.

1. Introduction

Diabetes affects a huge number of people all over the world. When beta cells inside the organ called pancreas is destroyed then generation of insulin from it is stopped. Inside the human body insulin provides signals to some carriers to be activated and carry the sugar from blood to cell-body. Insulin also sends signal to liver to store excess sugar and liver will use it later. When this chain is broken then sugar becomes captive inside blood and lack of sugar the cell cannot generate energy to drive the human body properly. The reasons of diabetes are: lack of exercise, living style, reaching at older age, genetically carrying the characteristics, taking excess carbohydrate, high blood pressure, and other factors. Diabetes can cause various serious diseases therefore detecting diabetes early can lower the patient's health risk. In this paper, we concentrated our efforts to detect the beginning of diabetes using several Machine Learning (ML) algorithms under Artificial Intelligence (AI).

Data clustering algorithms based on machine learning have recently been employed in real-world data to categorize them based on the outcome. Intense research is going on in the field both image and tabular data classification with the aid of machine learning algorithms. Few of previous works are discussion in this section. A comparative study of

supervised, unsupervised, reinforcement and semi-supervised learning techniques in data classification is found in [1].

Three types of classifications: binary or two category classification, Multiclass or multiple type classification and Multi-label classification are covered against eight different types of machine learning techniques. Similar concept is available in [2] for four ML algorithms, where authors work on five datasets to compare relative performance of four algorithms. In [3], authors compared five ML algorithms against Heart and Hepatitis disease accuracy, where Logistic Regression (LR) and Random Forest (RF) outperforms others. The clustering, classification, anomaly detection, deviation detection etc. are the basic approaches of data analysis. The application of clustering is not only in classification of data points but also acquiring distinctive features of engineering data or objects found in [4, 5]. Analyzing diabetic data is challenging since the majority of medical information is nonlinear, non- normal, and complex. Using machine learning approaches, Maniruzzaman, Md, et al. developed a model that identified and predicted Diabetes explained in [6]. Only the classification and the supervised system are addressed in that proposed method. They forecast diabetes mellitus using the PIDD (Pima Indian Dataset).

Machine learning developments have a substantial impact on the development of early diabetes diagnosis systems. Al-Tarawneh et al. proposed a model to identify diabetes using different machine learning algorithms named logistic Regression, Classification, Regression Tree, *K*-Nearest Neighbors and Multiple Perceptron algorithm. Authors created dataset, then ML algorithms was applied as effective instruments for diabetes identification found in [7].

Another study on detection of diabetes is found in [8], where the growing region technique is used in a segmentation strategy. Authors used 224 images of retinopathy eyes to test their method. The features are combined using fuzzy C-means and genetic algorithm for classification of patients. Another work in [9], which was built on diabetes issues, and the considered algorithm was Support Vector Machine (SVM). The authors were able to prove that their integrated approach could classify 99.4% of patents successfully. In [10] four MLs: LR, NB, SVM and RF are applied on diagnosis of diabetes using the dataset collected from Shalinitai Meghe Hospital and Research Centre, Nagpur, India. Same techniques are applied in [11] to detect diabetes, where the individual method found the accuracy below 80%. Comparison of five ML algorithms pertinent to detection of diabetes is shown in [12] using dataset in tabular form having 8 features. The results of five techniques are shown in bar graph, where accuracy of detection is also found below 80% for all the cases. Similar concept is available in [13] but the accuracy of detection is found above 90%.

In this paper, we used four data clustering algorithms to classify person having diabetes or not. Among the used algorithms, two unsupervised methods are: FCM, *K*-mean clustering and two supervised methods are: FIS and SVM. Here dimension of input vector is reduced by MLR and then normalized by Logistic Regression. None of the research papers mentioned above deal with four algorithms aided by MLR and combine them with entropy-based algorithm. This paper combined all the techniques with a simplified entropy based combining scheme to improve the accuracy of detection of diabetes.

The rest of paper is organized as: section 2 deals with some recent relevant works, section 3 provides basic theory of data clustering algorithm, section 4 provides steps of methodology along with complete description of fields of data set, section 5 provides results based on analysis of previous two sections: 3 and 4 and finally section 6 concludes the entire analysis.

2. Related Works

Usually, diagnosis of some common diseases like: diabetes, heart disease, functionality of liver or kidney etc. is obtained from details of medical report a patient. This job is done manually by a physician but when a large number of parameters are combined to come up with the solution then machine learning will help a doctor to take the correct decision. This section provides some previous works relevant to the paper. In [14] authors used 13 features of a patient then applied in ML to recognize heart diseases. Similar analysis is found in [15] under deep learning method. Sometimes two or more diseases are related in a sense that presence of one disease invites another one. In this context, the correlation between heart disease and Parkinson's Disease (PD) is detected in [16] using ML called Extreme Gradient Boosting. Diabetes detection is found under both ML and DL in [17], where accuracy of ML is found 81.3% and that of DL provides 89.07%. Three well-known MLs: Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) are used, where maximum accuracy is found for RF and minimum of LR. Three DL algorithms: Deep neural networks (DNN), Multilayer Perceptron (MLP) and Convolutional neural networks (CNN) are used and DNN provides the maximum accuracy. About ten ML algorithms are applied in [18] to detect diabetes using the data set with same attributes of this paper. Four parameters: Precision, Recall, F1 score are Accuracy are compared against all the algorithms but the main drawback of the paper is that algorithms are not combined to enhance accuracy of detection.

3. Basic Theory

In this paper we consider: FIS, FCM, *K*-mean clustering, SVM, MLR + logistic Regression for binary classification. This section provides basic of these ML techniques.

3.1. Fuzzy Inference System

The fuzzy inference system takes a set of linguistic variables each having crisp value as the input like fig.1. The

fuzzification unit converts the linguistic variable (with crisp value) to linguistic value. For example, very cold, cold, cool, warm, hot etc. are the linguistic values against the variable temperature. The probability of grade of each linguistic value is represented by membership function (MF), where the independent variable is called base variable like 0C (degree Celsius) for temperature. Next, the linguistic values under different Fuzzy variables are related by Fuzzy rules using the "IF THEN" rules along with connectors "AND" or "OR". The Fuzzy output is again converted to crisp value using defuzzification.

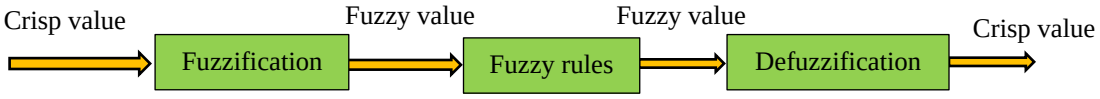


Fig.1 Fuzzy inference system

There are almost eight methods of defuzzification in recent literatures are reviewed and in this paper centroid method of de-fuzzification is used.

3.2. FCM and K-mean Clustering

FCM clustering and *K*-mean clustering both are unsupervised learning technique since they deal; with unlabeled data points. The property of a data point for example its distance from the center of a cluster determines its belonging to that cluster. In FCM this belonging is determined by the degree of MF of the point against that cluster, whereas *K*-means clustering is a direct approach found in [19-21].

3.3. Support Vector Machine (SVM)

SVM is a supervised learning technique uses labeled data set as: $\{(x_1, d_1), (x_2, d_1), (x_3, d_3), \ldots \ldots \ldots, (x_N, d_N)\}$, where x_i is the input vector (we call input pattern) and d_i is the desired response. Here a straight line called hyperplane is designed as,

$$L \equiv g(x) = w^T x + b = 0 \tag{1}$$

Sometimes curve of higher order polynomial is used instead of eq. (1), which separates the training samples, (x_i, d_i) in two regions keeping possible wide margin. The basic concept is: if a sample (x_i, d_i) provides positive value of $g(x)$ we consider $d_i = +1$ i.e. the sample is under class 1 and if $g(x)$ is negative then we consider $d_i = -1$ i.e. the sample point is under class 2 as found in [22, 23].

3.4. MLR and Logistic Regression

In MLR the input data X is a matrix of $M \times N$, where N is the dimension of data point vector and M is the number of data point vector. If the response vector $\mathbf{y}$ is $M \times 1$ then the coefficient vector is expressed as,

$$\alpha = (X'.X)^{-1}X'.y = [a_0\ a_1 \ldots a_{N-1}] \tag{2}$$

and the expression of MLR is,

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_{N-1}x_{N-1}. \tag{3}$$

To keep the outlier within [0 1] we use the exponential equation instead of linear equation like,

$$z = \frac{1}{1+e^{-(a_0+a_1x_1+a_2x_2+\ldots+a_{N-1}x_{N-1})}} \tag{4}$$

The eq. (4) is known as logistic regression as found in [24, 25].

4. Methodology

This section provides the combining scheme of results of four ML algorithms of section 3. The steps of works to deals with the numerical values of attributes are shown here explicitly. Finally, all the attributes are introduced with dataflow diagram.

4.1. Combining Scheme

The results obtained from four different MLs as mentioned in section 3 are combine by entropy-based technique of [26], but the algorithm is modified here to reduce complexity. If the accuracy of diabetes detection from *i*th ML is $p_i \leq 1$ then the entropy from the results of four MLs is, $H = \sum_{i=1}^{4} p_i \log_2\left(\frac{1}{p_i}\right)$. When $p_1 = p_2 = p_3 = p_4 = 1$ then $H = 0$ again

when $p_1 = p_2 = p_3 = p_4 = 0$ then $H = 0$. Therefore, entropy is 0 for both the best and worst cases. Entropy is maximum when $p_1 = p_2 = p_3 = p_4 = 0.25$ and the maximum value is $H = 2$. Under above three condition the combine accuracy of detection is made using the following technique.

- Store the accuracy of detection $[p_1, p_2, p_3, p_4]$
- Normalize the probability of detection dividing the individual by the sum as: $p_{ni} = p_i/(p_1 + p_2 + p_3 + p_4)$
- Evaluate the entropy, $H = \sum_{i=1}^{4} p_{ni}\, log_2\left(\frac{1}{p_{ni}}\right)$
- If $H < 0.15$ and at least two probability of detection is greater than 0.75 then detection is considered correct; Else, imperfect detection of diabetes.

4.2. Steps of Algorithm

The methodology used in the paper is given as:

Step 1: select the data table
Step 2: apply MLR to reduce dimension of data.
Step 3: apply logistic regression on data to keep outlier within [0 1]
Step 4: apply FIS on the tabular data and determine graphical representation of fuzzy rules.
Step 5: apply *K*-mean clustering and FCM to cluster the preprocessed data under MLR and logistic regression.
Step 6: use SVM to segregate data with optimum space
Step 7: determine accuracy of detection of all the methods
Step 8: calculate accuracy of all methods using entropy based combining scheme of sub-section 4.1.

4.3. Dataset Used and Flow Diagram

Pima Diabetes dataset (https://www.kaggle.com/uciml/pima- indians-diabetes-database) is used in this paper, taking 768 records each with 9 attributes. Few data in tabular form are shown in appendix. The attributes are introduced as:

- PREG: the number of pregnancies a woman has.
- PLAS: It is the plasma glucose concentration.
- PRES: Blood pressure in the patient's arm (diastolic)
- SKIN: The thickness of a patient's skin is measured at the triceps.
- INSULIN: amount of insulin (2 hrs serum)
- BMI: stands for body mass index, which is a ratio of *w* to *h*. *w* = weight and *h* = height
- PEDI: In diabetes prognosis, an appealing attribute is used.
- Age: the age of patients
- Outcome: 'Yes' indicates that the patient is diabetic, whereas 'No' indicates that the patient is not diabetic.

The dataflow model proposed in the paper is given in fig.2.

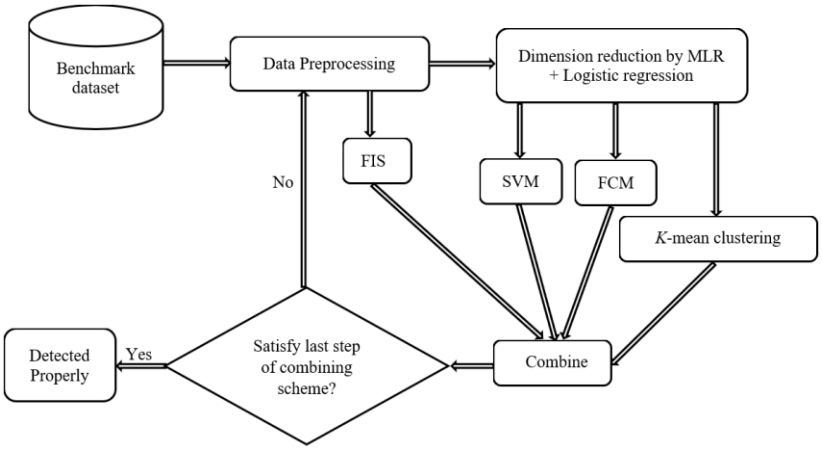


Fig.2. Flowchart of proposed model

The proposed methodology will evaluate the performance of individual ML algorithm and will also examine the any improvement of accuracy with their combination.

5. Results

The data of diabetes used in this research work has eight fields: Pregnancies, Glucose, Blood Pressure, Skin

Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. The input data is found uncorrelated against the outcome of 'diabetes' and 'no diabetes'. First of all, we apply combination of MLR and logistic regression on the dataset. The MLR reduces the dimension of data and logistic regression confines data points within [0 1]. The output of regression against the index value as shown in fig.3. Here we consider 84 records of data table among them first 42 is the case of 'no diabetes' (outcome is indicated by 0) and the next 42 records are taken for the case of diabetes and corresponding output is labeled as 1. If the threshold level is taken 0.5 then it is visualized that MLR logistic regression is able to classify the data with very little error.

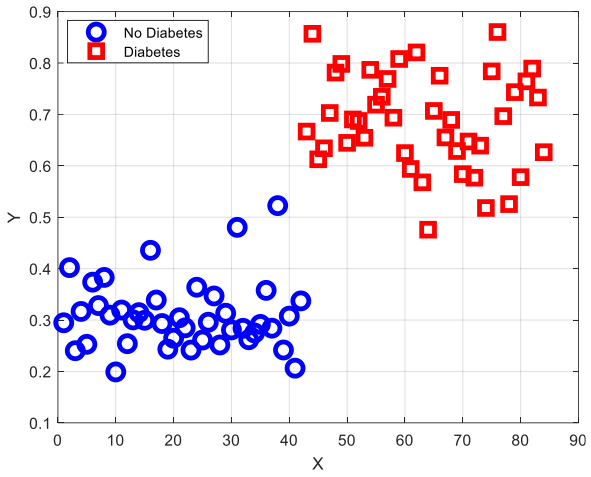
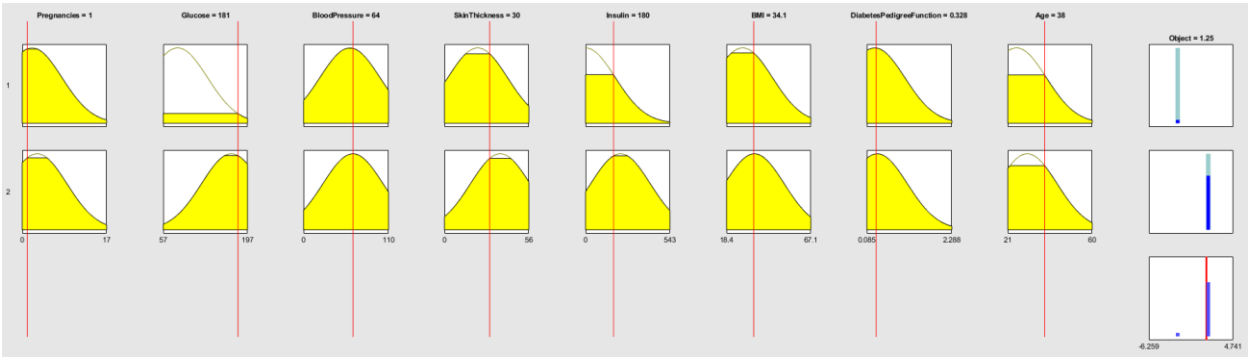
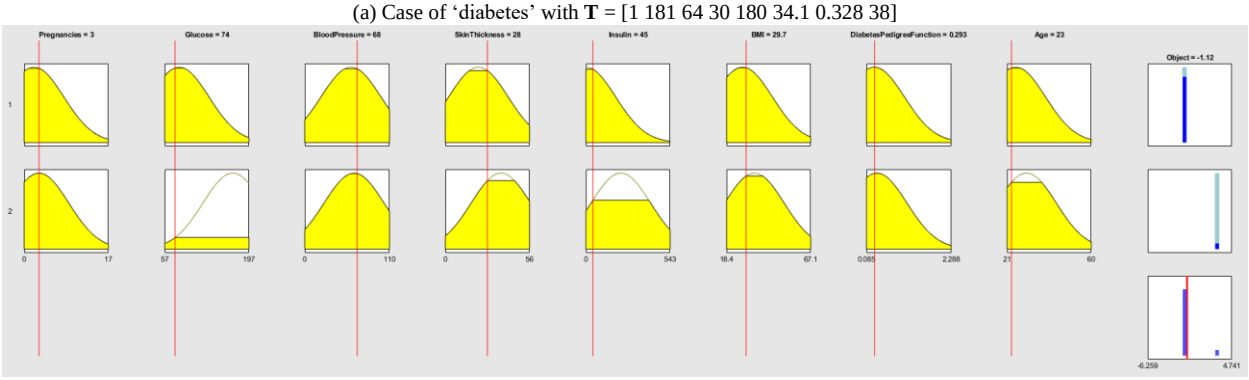


Fig.3. MLR and Logistic regression on the data

Next, we applied the data vectors (or records of table) on the FIS of eight inputs where each input corresponds to an attribute of the data table and output of the FIS is binary i.e. +1 for the case of 'diabetes' and -1 for the case of 'no diabetes'.



(a) Case of 'diabetes' with **T** = [1 181 64 30 180 34.1 0.328 38]



(b) Case of 'no diabetes' with **T** = [3 74 68 28 45 29.7 0.293 23]

Fig.4. Verification of fuzzy rules of FIS

The FIS is trained with 500 data then the FIS rule is verified by the following two test records as an example:

$\mathbf{T}_1$ = [1 181 64 30 180 34.1 0.328 38]
$\mathbf{T}_2$ = [3 74 68 28 45 29.7 0.293 23]

Where first one is the case of ‘diabetes’ and the second one is the case of ‘no diabetes’. The test results are shown in fig.4 (a) and (b) respectively under Fuzzy rules of FIS. Taking threshold value $\tau = 0$, we can segregate the results of pathological reports. The first input vector gives the output of 1.25 (positive) and the second input vector provides the output of -1.12 (negative). In FIS we normalized the data set dividing them by 10. The variation of normalized variables are shown by surface plot under FIS in fig.5. Some variables show linear relation and some of them show non-linear relation with the fuzzy output.

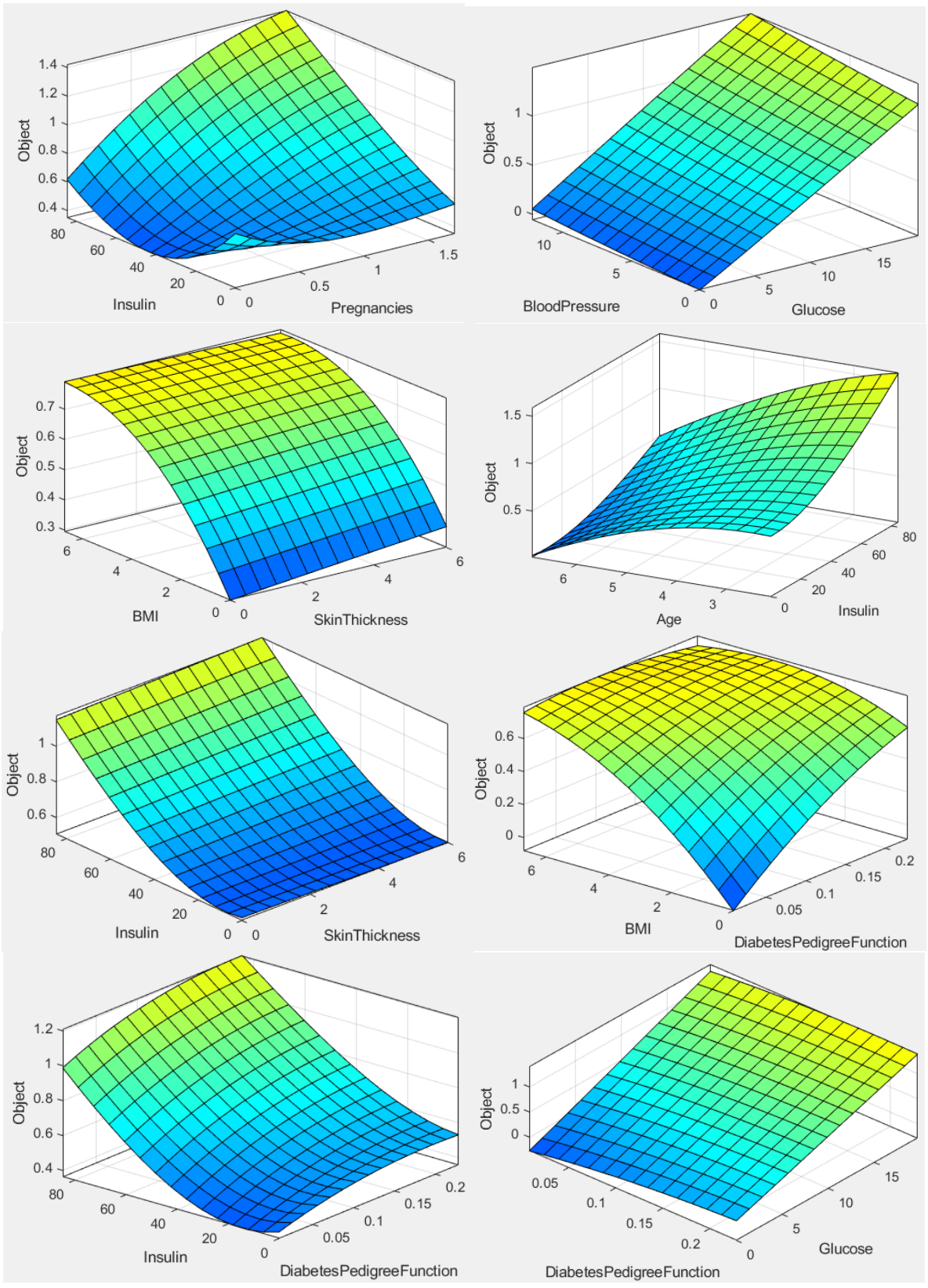
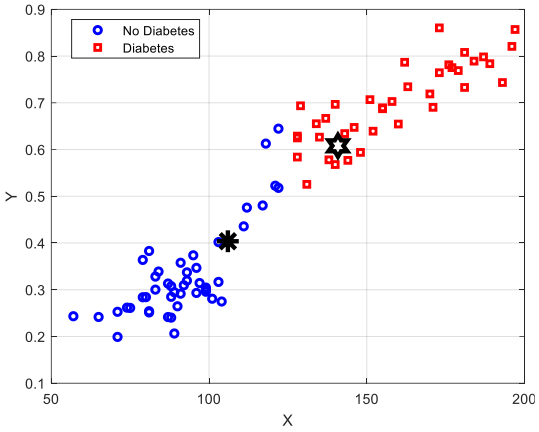
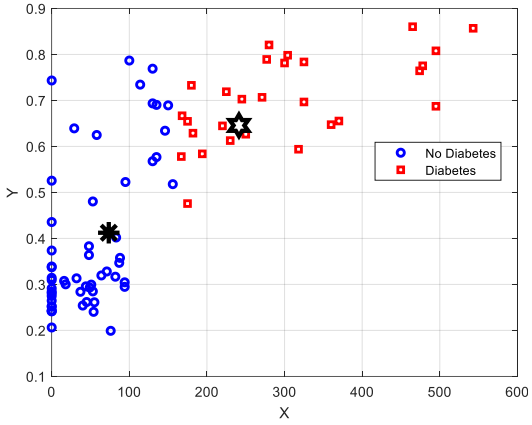


Fig.5. Surface plot of FIS

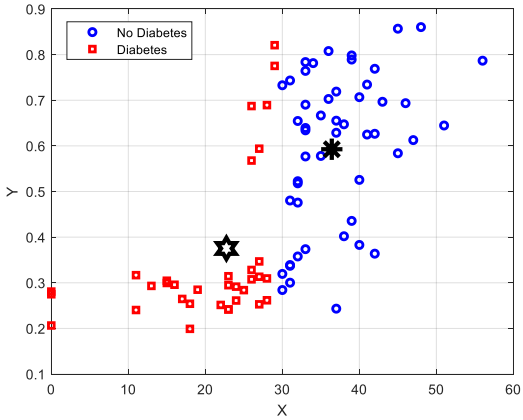
Fig.6 shows five scatterplots under different combination of fields of data table using FCM. Here 84 data points are plotted, where 42 for the case positive report and 42 for the negative report. Before applying FCM, we take the output of combination of MLR and logistic regression on eight variables. The normalized output of MLR + Logistics is taken along *y*-axis and different fields (or attributes) of data table are taken along *x*-axis. In this case, the scattered data points are separable along both horizontal axis (along the data variable) and also along vertical axis (output of regression under 8 variables). The result of fig.6 indicates medium correlation of data under three fields: Glucose, Insulin and Skin thickness shown in fig.6 (a)-(c). The other two variable BMI and age not correlated with outcome directly as found in fig.6 (d)-(e). Similar results are shown in fig. 7(a)-(e) for *K*-mean clustering and results are almost similar like FCM. Therefore, both FCM and *K*-mean clustering reveal very poor classification. The big stars on the scattered plot indicate the center of each class.



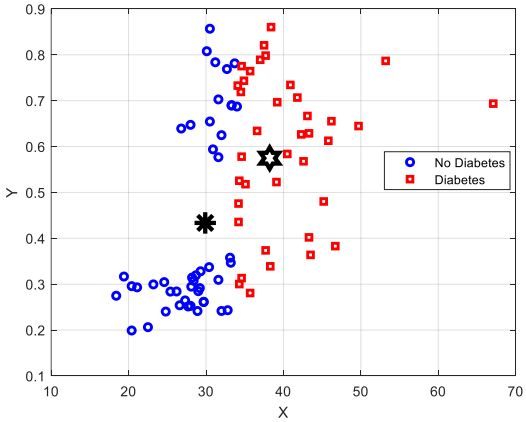
(a) MLR vs. Glucose



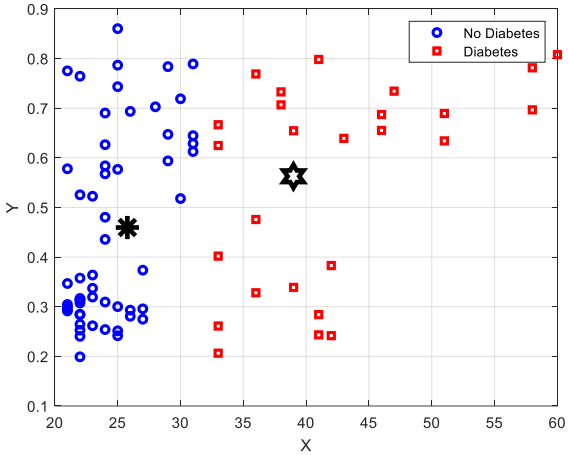
(b) MLR vs. Insulin



(c)MLR vs. Skin thickness



(d) MLR vs. BMI

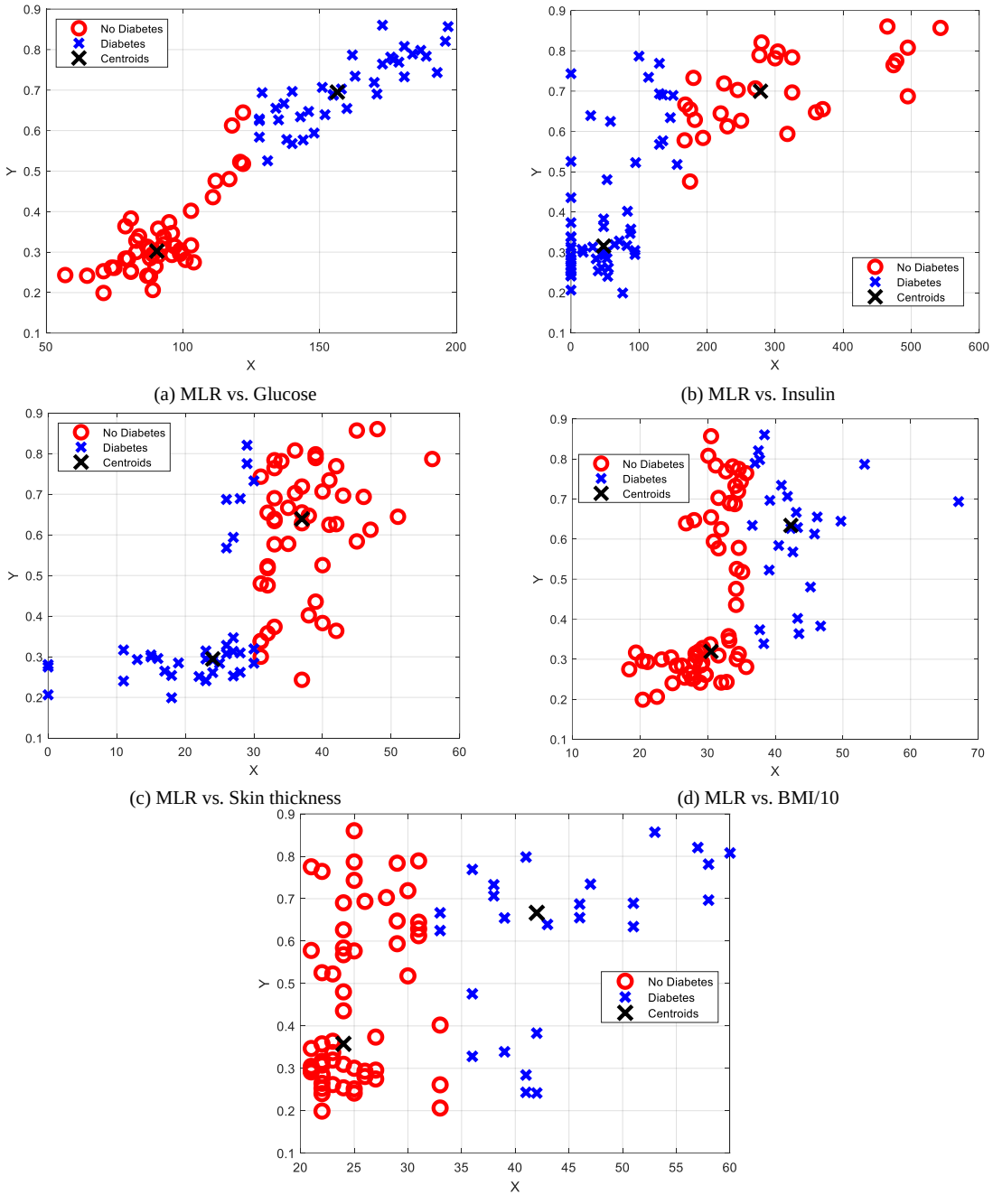


(e) MLR vs. age

Fig.6. Scatterplot of FCM clustering

Fig.8 (a)-(d) shows the results for combination of MLR and SVM, where *y*-axis is the output of regression and *x*-axis is the 'attributes of table' for first three graphs and 'index of data' for the fourth graph. Here we consider four most

important parameter: Glucose, Skin thickness, Insulin and 'index of regression' along *x*-axis of the graphs. It is visualized that a few points fall on wrong side of hyper-plane except the fourth graph. The variation of objective function against the number of iterations are shown in fig.9 (a)-(d) considering Glucose, Skin thickness and Insulin gain as the attributes. Among the four cases, 'Index of regression' reveals the best result as found in fig. 8(d) and fig. 9(d).



(a) MLR vs. Glucose

(b) MLR vs. Insulin

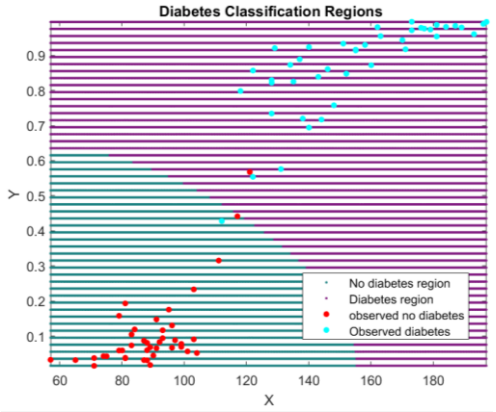
(c) MLR vs. Skin thickness

(d) MLR vs. BMI/10

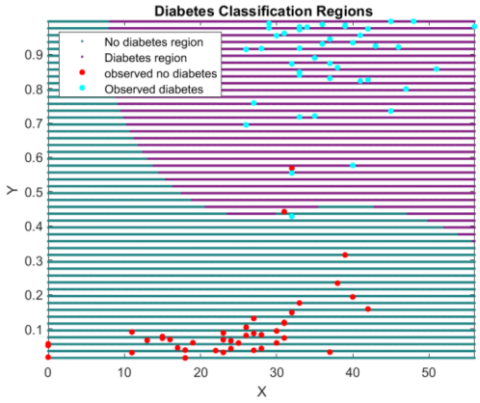
(e) MLR vs. Age

Fig.7. Scatterplot of K-mean clustering

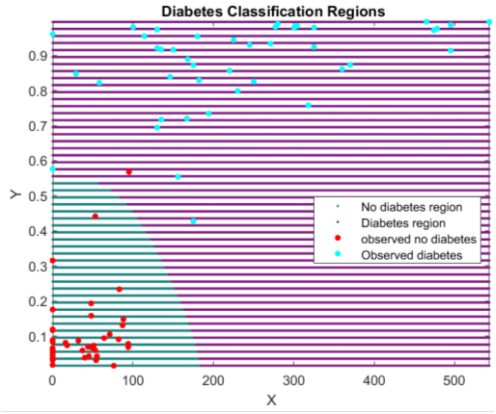
The parameter 'box constraint' governs the margin of SVM, hence has the direct impact on outlier. If box constraint is increased it reduces the margin and SVM provides fewer points of support vectors. In this case training time will be longer. For wider margin of hyper plane, the separation of classes is more distinct. Now, process time will be low at the expense of outlier points (it is increased), therefore box constraint needs to be optimized. There is a tradeoff between width of margin and errors of training data. In SVM the scale parameter (for example γ parameter of Gaussian RBF) of the kernel function needs to be scaled properly before applying the kernel function on the data points. The performance of SVM also depends on scale parameter. The optimum value of scale parameter and box constraint minimizes the objective function. The 3D plot of kernel scale, box constraint, objective function for the same parameters is shown in fig. 10(a)-(d) against four independent variables. Comparison of parameters of SVM against four fields are shown in Table 1 to get the impact attributes on the parameters of SVM.



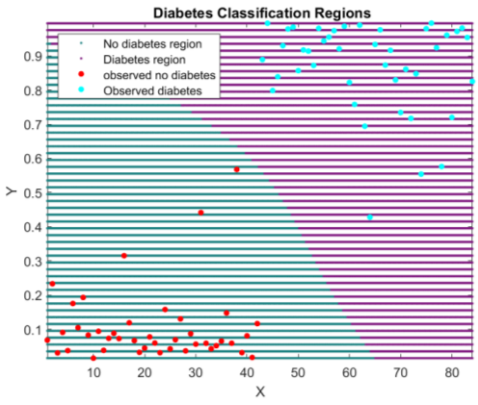
(a) MLR vs. Glucose



(b) MLR vs. Skin thickness

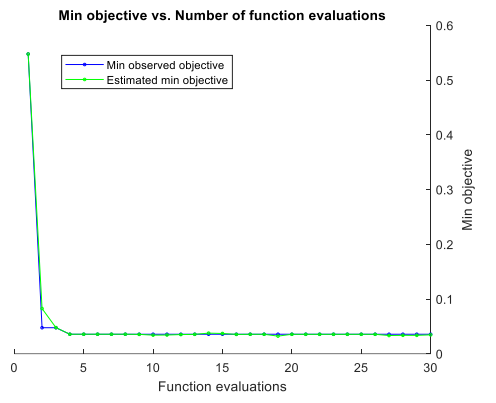


(c) MLR vs. Insulin

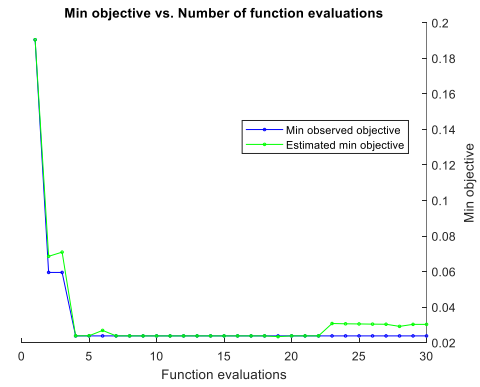


(d) MLR vs. Index of regression

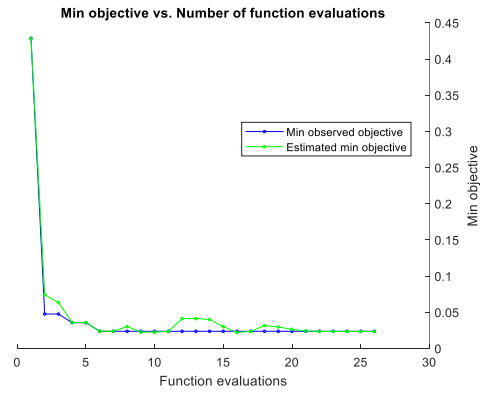
Fig.8. Colored region with data under support vector machine



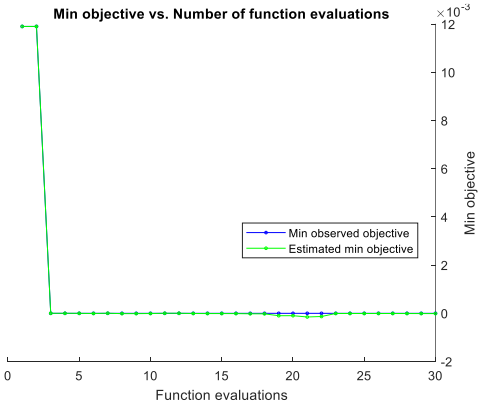
(a) Glucose



(b) Skin thickness

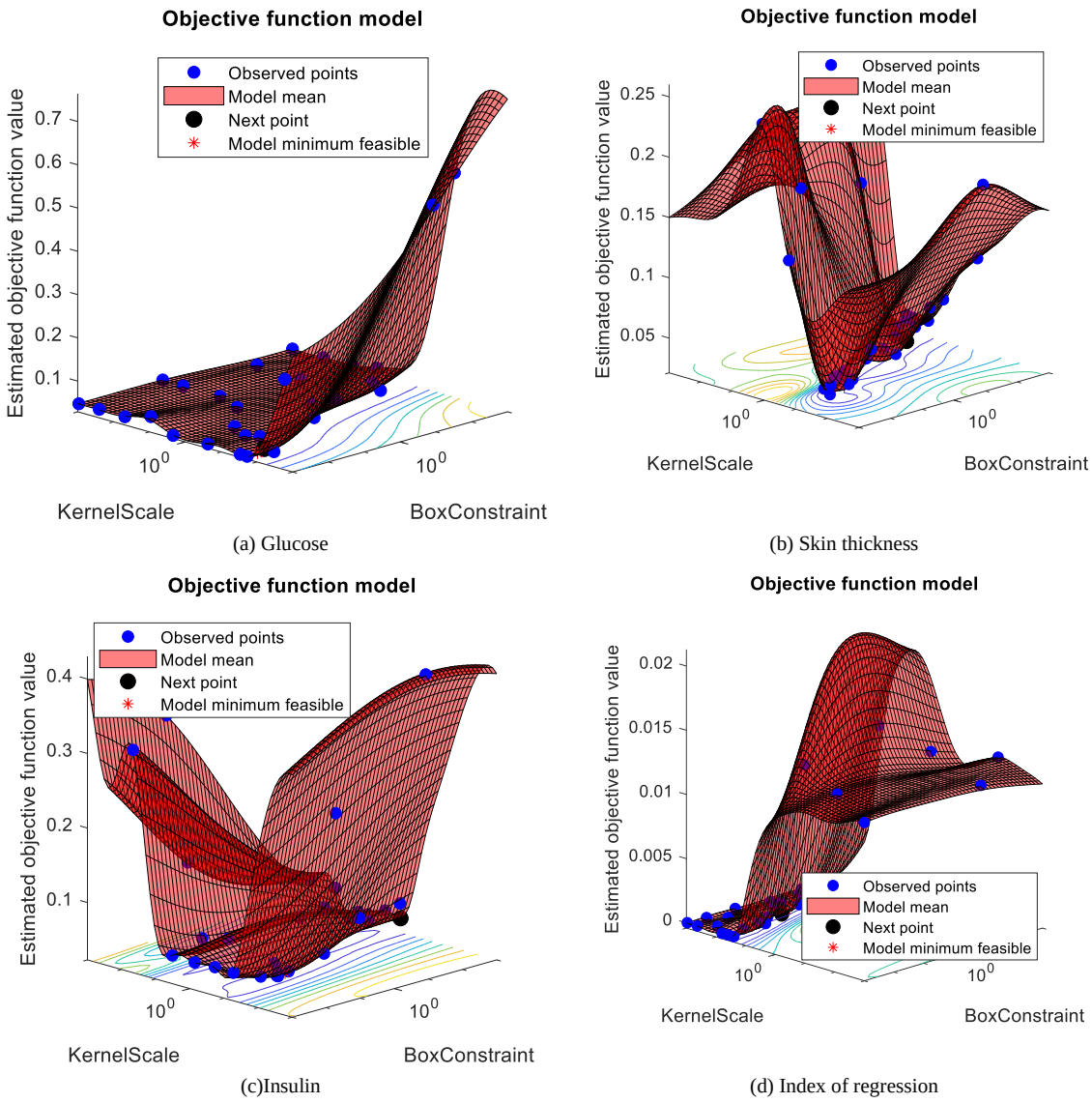


(c) Insulin



(d) Index of regression

Fig.9. Variation of objective function against the number of iterations



(a) Glucose (b) Skin thickness

(c)Insulin (d) Index of regression

Fig.10. Variation of 'objective function' on 'kernel scale' and 'box constraint'

Table 1. Comparison of SVM parameters

Fields	Outlier Rate	Box Constraint	Kernel Scale	Elapsed time (sec)
Glucose	0.0833	0.014824	0.12256	30.493
Skin thickness	0.1071	0.019943	0.12545	26.6057
Insulin	0.0952	0.20643	0.26245	50.1943
Index of regression	0.0595	0.22938	261.18	19.7881

Table 2. Accuracy of detection

Algorithms	Glucose	Skin thickness	Insulin	Index of regression	Combined method
MLR + Logistic	0.821	0.743	0.792	0.844	0.9413
FIS	Considering all fields: 0.894				
FCM	0.739	0.711	0.721	0.784	
K-mean Clustering	0.744	0.716	0.705	0.726	
SVM	0.735	0.702	0.681	0.773	

If points are trained by MLR + Logistic regression then application of SVM on 'index' vs. 'output of regression' shows some better results visualized from fig. 8 and fig. 9. Finally, accuracy of detection of individual method and that of derived from a modified version of entropy based combined model of [26] is shown in Table 2. The modified version

of entropy based combination technique is shown in sub-section 4.1 of the paper. The combined method gives the highest accuracy in identification of diabetes, which is above 94% and better than individual method.

6. Conclusions

In this paper four data clustering algorithms: FCM, *K*-means clustering, FIS, and SVM are used both individually and in combined form to detect the presence of diabetics. In our real life, the physicians usually observe the medical reports of a patient and takes the decision for intuitive point of view. When large number reports with numerical and linguistic values are accumulated together to take the decision then there may have a chance of erroneous decision. In this case, the model of the paper can help a physician to verify his experience. The methodology used in paper can be used in any binary classification irrespective of dimension of input vector, since dimension reduction is done by MLR and logistic regression. The combined method provides the accuracy of detection above 94%. In the future, we will consider deep leaning techniques like CNN, RNN, LSTM and will observe the improvement of classification on the same data along with a comparison will be made about process time.

References

[1] Iqbal, H. S., Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, vol. 2, no. 160, pp.1-12, 2021.

[2] Vraj, S., Urvashi, T. and Ankit S., A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. 4th International Conference on Innovative Data Communication Technology and Application, Elsevier, vol. 215, pp. 422-431, 2022.

[3] Ul Hassan, C., Khan, M. and Shah, M., Comparison of Machine Learning Algorithms in Data classification. 2018 24th International Conference on Automation and Computing (ICAC), Newcastle Upon Tyne, UK, pp. 1-6, 2018.

[4] Danyang, C. and Bingru, Y., An improved k-medoids clustering algorithm. In 2010 2nd International Conference on Computer and Automation Engineering (ICCAE), Singapore, pp.132-135, 2010.

[5] Harpreet, S., Madan, M., Thomas, M., Zeng-Guang, H., Kum Kum, G., Ashu, S. and Lotfi Z. Real-life applications of fuzzy logic. Hindawi Publishing Corporation Advances in Fuzzy Systems, vol. 2013, Article ID 581879, pp.1-3, 2013.

[6] Maniruzzaman, M., Rahman, M., Ahammed, B. and Abedin M., Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems, vol. 8(1), pp. 1-14, 2020.

[7] Al-Tarawneh, M., Mustafa, M. and Al-Tarawneh, Z., Hand Movement-Based Diabetes Detection Using Machine Learning Techniques. vol. 9, ISSN 2281-2881, pp. 1-13, 2021.

[8] Ghoushchi, S., Ranjbarzadeh, R., Dadkhah, A., Pourasad, Y. and Bendechache, M., An extended approach to predict retinopathy in diabetic patients using the genetic algorithm and fuzzy C-means. BioMed Research International, vol. 2021, pp. 1-13, 2021.

[9] Devi, R., Bai, A. and Nagarajan N., A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. Obesity Medicine, Elsevier, vol. 17, 2020.

[10] Dubey, Y., Wankhede, P., Borkar, T., Borkar, A. and Mitra, K., Diabetes Prediction and Classification using Machine Learning Algorithms. 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, pp. 60-63, 2021.

[11] Islam, N. and Khanam, R., Classification of Diabetes using Machine Learning. International Conference on Computational Performance Evaluation (ComPE), Shillong, India, pp. 185-189, 2021.

[12] Wei, S., Zhao, X. and Miao, C., A comprehensive exploration to the machine learning techniques for diabetes identification. IEEE 4th World Forum on Internet of Things (WF-IoT), Singapore, pp. 291-295, 2018.

[13] Dalve, P., Bobby, D., Marathe, A., Dusane, A. and Daga, S., Comparison of Performance of Machine Learning Algorithms for Diabetes Detection. Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, pp. 1-7, 2023.

[14] Karthik, K., Reddy, A., Kulkarni, R. and Mehdi, M., Algorithm Accuracy Verification in Heart Disease Analysis using Machine Learning. 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, pp. 345-349, 2023.

[15] V, C. and Baby, S., Systematic Review on Deep Learning-based Heart Disease Diagnosis. 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, pp. 908-912, 2023.

[16] Kumar, N., Avasthi, S. and Prakash, A., Establishing the Correlation between Parkinson's and Heart Disease using Machine Learning Algorithm. International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, pp. 581-586, 2023.

[17] Boon, W., Saaveethya, S., King Hann, L., Wong, W. and Filbert, H., Diabetes detection based on machine learning and deep learning approaches. Multimedia Tools and Applications, pp.1-33, 2023.

[18] Shahin A., Khairul, I., A. Arjan, D., Duranta, D., Farija, H. and Habibur, R., A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights. vol. 2023, pp.1-15, 2023.

[19] Handoyo, M. and Imam, P., The Fuzzy Inference System with Rule Bases Generated by using the Fuzzy C-Means to Predict Regional Minimum Wage in Indonesia. International J. of Opers. and Quant. Management, vol. 24, no. 4, pp. 277-292, 2018.

[20] Memon, K. and Lee, D., Generalized kernel weighted fuzzy C-means clustering algorithm with local information. Fuzzy Sets and Systems, vol. 340, pp. 91-108, 2018.

[21] Rahmani, M., Pal, N. and Arora, K., Clustering of image data using K-means and fuzzy K-means. International Journal of Advanced Computer Science and Applications, vol. 5, pp. 160-163, 2014.

[22] Özaltın, Ö. and Yeniay, Ö., ECG Classification Performing Feature Extraction Automatically Using a Hybrid CNN-SVM Algorithm. 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, Ankara, Turkey, pp. 1-5.

[23] Wang, Y. and Wu, Q., Research on Face Recognition Technology Based on PCA and SVM. 7th International Conference on Big Data Analytics (ICBDA), Guangzhou, China, pp. 248-252, 2022.
[24] Tigga O., Pal J. and Mustafi D., A Comparative Study of Multiple Linear Regression and K Nearest Neighbours using Machine Learning. Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, pp. 1-5, 2023.
[25] Gufran, A. A., Salliah, S. B., Mohd, D. A., Sultan, A., Jabeen, N. and Eljialy, A., Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction. Computational and Mathematical Methods in Medicine Volume 2023, Article ID 8191261, pp.1-10.
[26] Kalam A., Anup, M., Juga, D. and Imdadul I., Improving signal detection accuracy at FC of a CRN using machine learning and fuzzy rules. Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 2, pp. 1140-1150, 2021.

Appendix

Table 3. Some Training Data Records

SL	Pregnancies Woman	Plasma Glucose	Blood Pressure	Skin Thickness	Insulin	Body Mass	PEDI	Age	Report
1	13	145	82	19	110	22.2	0.245	57	Negative
2	13	106	72	54	0	36.6	0.178	45	Negative
3	13	106	70	0	0	34.2	0.251	52	Negative
4	13	76	60	0	0	32.8	0.18	41	Negative
5	13	153	88	37	140	40.6	1.174	39	Negative
6	12	106	80	0	0	23.6	0.137	44	Negative
7	14	175	62	30	0	33.6	0.212	38	Positive
8	13	126	90	0	0	43.4	0.583	42	Positive
9	13	152	90	33	29	26.8	0.731	43	Positive
10	13	129	0	30	0	39.9	0.569	44	Positive
11	13	104	72	0	0	31.2	0.465	38	Positive
12	13	158	114	0	0	42.3	0.257	44	Positive

Authors' Profiles

Shifat Jahan Setu completed her M.Sc. from the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh in 2021. She also obtained her B.Sc. degree from the same university in 2019. Currently she is working as a Lecture at the department of Computer Science & Engineering in Dhaka International University (DIU), Dhaka, Bangladesh. Her research interests are Machine Learning, Data Mining and Artificial Intelligence.

Fahima Tabassum is currently serving as a Professor at Institute of Information Technology, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh. She has completed her B.Sc. (Hons.) from the department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh in 2003 and M.Sc from the same department in 2010. She has achieved the degree of Doctor of Philosophy in Image Processing from the same department in 2023. She is currently doing his research in image processing, machine learning and software system analysis and development.

Sarwar Jahan is serving as an Associate Professor in the Department of Computer Science and Engineering at East West University, Dhaka, Bangladesh. He received his B.Sc. degree in Electrical and Electronics Engineering from Ahsanullah University of Science and Technology, Dhaka, Bangladesh, and M.S. degrees in Telecommunication Engineering from the University of Technology, Sydney, Australia in 2001, and 2005 respectively. He has completed his Ph.D. degree from the Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh in the field of wireless communications in 2022. He is currently doing his research in Communication Engineering, Network Traffic, and different disease detection using artificial intelligence and machine learning algorithm.

Md. Imdadul Islam has completed his B.Sc. and M.Sc Engineering in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh in 1993 and 1998 respectively and has completed his Ph.D degree from the Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh in the field of network traffic in 2010. He is now working as a Professor at the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. Previously, he worked as an Assistant Engineer in Sheba Telecom (Pvt.) LTD (A joint venture company between Bangladesh and Malaysia, for Mobile cellular and WLL), from Sept.1994 to July 1996. Dr Islam has a very good field experience in installation and design of mobile cellular network, Radio Base Stations and Switching Centers for both mobile and WLL. His research field is network traffic, wireless communications, wavelet transform, adaptive filter theory, ANFIS, neural network, deep learning and machine learning. He has more than two hundred research papers in national and international journals and conference proceedings.

How to cite this paper: Shifat Jahan Setu, Fahima Tabassum, Sarwar Jahan, Md. Imdadul Islam, "Detection of Diabetes using Combined ML Algorithm", International Journal of Intelligent Systems and Applications(IJISA), Vol.16, No.1, pp.11-23, 2024. DOI:10.5815/ijisa.2024.01.02