Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection

Sajratul Yakin Rubaiat
Computer Science and Engineering Department
Patuakhali Science and Technology University
Patuakhali, Bangladesh
Email: yakin13@cse.pstu.ac.bd

Md Monibor Rahman
Electrical and Computer
Engineering Department
Old Dominion University,USA
Email: mrahm006@odu.edu

Md.Kamrul Hasan
Mathematics, Statistics and
Computer Science Department
Marquette University, USA
Email: mdkamrul.hasan@mu.edu

***Abstract*—More than 400 million people in the world have diabetes. High-risk factors of diabetic individuals vary dramatically, and many patients suffer complications and avoidable harm. Improving the identification level of high-risk factors would help to reduce the rate of complications. To do this, it is essential to analyze a person's medical record, detailed health information that currently requires doctors and is manual, time-consuming, and subjective. In this work, we introduce an approach to automatically predict type 2 diabetes mellitus (T2DM) applying a neural network. The objective of this paper is to find which type of model that works best for predicting diabetes. We used the Pima Indian Diabetes data-set in this analysis. The analysis was carried out on this database using two methods. The first method includes Data Recovery followed by feature selection. We input these features to the MLP neural network classifier which achieved an accuracy of 85.15%. In our second approach, we applied noise reduction based method using k-means followed by feature selection. The features thus obtained are used with Random Forest, Logistic Regression and MLP neural network classifier. The maximum accuracy obtained among these classifiers is 77.08%. The consultation shows why Data recovery with MLP is far better than K-means based noise reduction with the different type of classifier.**

I. Introduction

Diabetes Mellitus (DM) is a significant public health problem that is approaching epidemic proportions globally [1]. It has notably increased in the 21 century. Diabetes is caused by several factors, including obesity, consumption of unhealthy food, heredity, etc. As of 2015, about 415 million people have had diabetes worldwide, and the trend suggests that the rate will continue to rise. Diabetes has some serious long-term complications including cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes. For these reasons, researchers need to put more focus on this problem. There are three kinds of diabetes. First, Type 1 DM which is caused by a failure of the pancreas to produce sufficient insulin. Second, Type 2 that is the most common DM and the reason is identified as excessive body weight and insufficient exercise. Third, gestational diabetes which occurs in pregnant women with no prior history of diabetes. It is pointed out that Type 2 DM makes up about 90% of the cases.

Data analysis has been successfully applied to various fields of human society, such as weather prognosis, market analysis, engineering diagnosis, and customer relationship management. However, the utilization of disease prediction and medical data analysis still has room for improvement. Every hospital possesses a different kind of necessary and medical information, and it is essential to extract useful information from these data to support future medical analysis and diagnosis [2, 3]. It is rational to believe that there are several valuable patterns and are waiting for researchers to examine them.

As the number of the patient of diabetes patients is increasing, it is necessary to build a model that can classify patients with high risk of diabetes in the future. In the future, the identified high-risk factors could potentially prevent more cases of diabetes.

II. Pima Indian Diabetes Database

Pima Indian data-set [2] is obtained originally from the database of National Institute of Diabetes and Digestive and Kidney Diseases of the United States. The objective of the data-set is to diagnostically predict whether or not a patient has diabetes, based on specific diagnostic measurements included in the data-set. Several constraints were placed on the selection of these instances from a more extensive database. In particular, all patients here are females from Pima Indian Heritage who are at least 21 years old. The information consists of 768 patients (268 instances of 1 and 500 cases of 0) coming from a population near Phoenix, Arizona, USA. 1 and 0 indicates whether the patient has diabetes or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. The data-set consists of several medical predictor variables one target variable and the outcome. A number of predictor variables is available in the database, including number of times pregnant (preg), plasma glucose concentration at 2h in an oral glucose tolerance test (plas), diastolic blood pressure (pres), triceps skinfold thickness (skin), 2-h serum insulin (insu), body mass index (bmi), diabetes pedigree function (pedi), age, and class variable (class).

III. Related works

Artificial intelligence (AI) techniques are used today to enhance and improve our regular lifestyle. Use of AI techniques span modeling and analysis of hemoglobin level identification [3]–[7], activity detection [8], [9], pain level detection [10], [11], as well as prediction model to identify the high risk factors to prevent diabetic issues, including [12]–[15]. Kamer Kayaer et al. [12] used the PID dataset to evaluate the perceptron-like general regression neural network (GRNN). This study had 576 cases in the training set and 192 cases in the test set. Using 576 training instances, the sensitivity and specificity of their algorithm was 80.21% on the remaining 192 instances. The same number of random training and test sets was used to compare the simulation results. Dilip Kumar et al. [13] used naive Bayes with the genetic algorithm to evaluate the perception. The accuracy and specificity of their algorithm was 78.69% on the test set. Manjeevan S. et al. [14] used fuzzy min-max (FMM) neural network to evaluate the model and the accuracy of the algorithm was 78.39%. Hayashi, Y., & Yukita, S.(2016) [15] used Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques and get 83.83% accuracy.

IV. Methodology

We introduce a rich analysis of the features available in the PIMA database to identify the risk factor of a diabetic. We develop two approaches leveraging machine learning algorithm, which to our knowledge has not been previously studied. We show that our method is able to both efficiently detect the risk factors as well as significantly outperform previous works on (diabetic diseases) risk factor detection tool. Out first method is accomplished applying neural network model. The second process is developed based on the K-means algorithm.

A. Neural network-based method

We use a neural network model in this process. We have three different steps in this section such as data recovery, feature selection, and M.L.P. Classifier. As a first step, data recovery techniques are applied by replacing the missing data with the mean value for making the dataset complete for building a model. Then, we do the feature selection process which is done to find the features that have the most impact on the risk factor identification. Lastly, a suitable number of hyper-parameters are selected that works well for this data-set.

B. K-means-based method

We apply the K-means algorithm in this section after selecting the features. We also use different machine learning classifiers to compare the outcome. The k-means algorithm effectively reduces noise from the data. The output of the k-means algorithm is used as a feature for the model. The classification methods are applied to the selected features to see the result.

V. Applying data recovery with neural network model

In this process, we apply data processing, feature selection, and machine learning algorithm sequentially.

A. Data recovery

Pima Indian dataset contains various missing data in several features including blood pressure, insulin level, skin thickness, blood pressure, BMI, and glucose levels. We observe zero entries in 374 insulin, 227 skin thickness, 35 blood pressure, 11 BMI, and five glucose features since the missing data affect the data analysis process and mislead the prediction results of the model. Model built with this data would be misleading. There are different methods to preprocess the data. For example, we can delete the data from the data set, replace the missing data with the mean value, or replace the missing data with the most likely value of this feature.

Pima Indian dataset is minimal, only 768 samples. Therefore, the model can end up being highly biased if the training data are deleted. So, removing observations from the training set is not a good idea. Different person has a different level of insulin level. If we transfer the high number of data with the most likely value, then we may face the issue with high variance in the data. So, the best option is replacing the missing values with the mean value of that particular attribute. As a part of preprocessing, we replace the missing data with a `NaN` value at first. Here, `NaN` is used for replacing the numerical missing value with a string. Afterward, we iterate over the column and find the sum of all the numerical values (`NaN` is a string value, so it does not add up here). Then, we calculate the general mean by dividing total summation by the number of the entity in the column. Finally, we replace the `NaN` string with the numeric mean value.

B. Feature selection

All the extracted features do not carry significant weight, and some of them do not have any impact on the prediction model. Apply this kind of feature for training the model will only add up the additional computational power. Fig. 1 contains the Skin thickness graph of all the patient.

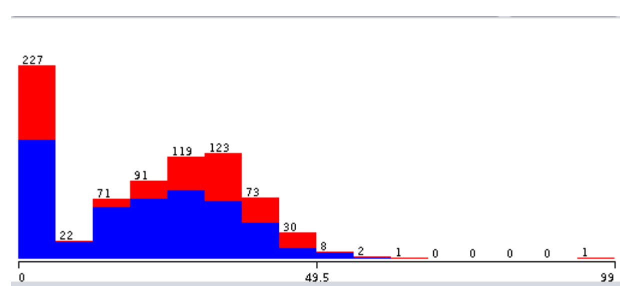

Fig. 1. Skin Thickness graph for "Pima Indian data set". Here X-axis contain skin thickness, Y-axis Contain Number of patient that have diabetes positive(as red) and diabetes negative(as blue).

The first block of the histogram contains a pretty much same number of element from both classes (diabetes positive or diabetes negative). And it still maintains this ratio when the skin thickness is increased. So this cannot be an essential feature for the model.

These data from Fig. 2 and Fig. 3 concludes that when glucose and B.M.I. levels increase, the risk for diabetes rises significantly. So it provides us with a useful linear property with the B.M.I or glucose level.

Greedy Stepwise Search Algorithm to select the critical attributes. Greedy Stepwise Search Algorithm iterates through each or set of the attribute to calculate which property gives the minimum error. The step for the feature selection algorithm is as follows:

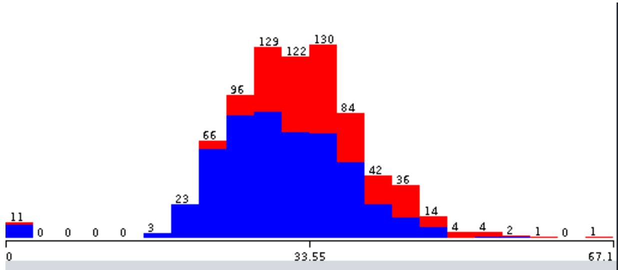

Fig. 2. B.M.I. feature graph from Pima Indian data set(Here X-axis contain B.M.I And Y-axis contain the Number of patient that have diabetes positive(as red) and diabetes negative(as blue).

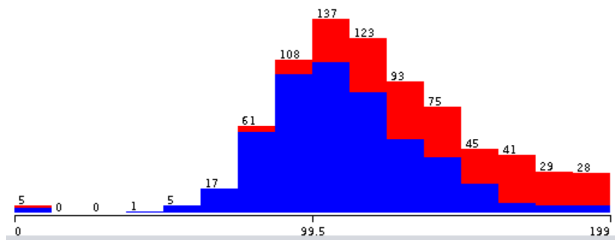

Fig. 3. Glucose level graph for Pima Indian data set(Here X-axis contain glucose level, Y-axis contain Number of patient that have diabetes positive(as red) and diabetes negative(as blue).

1) Pick a dictionary of feature h_0(x),...h_D(x)
   - e.g., polynomials for linear regression
2) Greedy heuristic:
   i Start with empty set of feature F_0 = Ø (or simple set, like just h_0(x) $\rightarrow$ y_i + ε)
   ii Fit model using current feature set F_i to get w^t
   iii Select next best feature h_j.(x)
      - e.g., h_j(x) resulting in lowest training error where learning with F_i+h_j(x)
   iv Set $F_t + 1 \Leftarrow F_t$ + h_j.(x)
   v Recurse

By analyzing the different graph and investing which feature affects the most, four features were taken. Those features are given below:

- Glucose
- B.M.I.
- Diabetes Pedigree Function
- Age

TABLE I
SAMPLE DATA RESULTING FROM APPLYING GREEDY STEP-WISE ALGORITHM

Glucose	BMI	Diabetes Pedigree Function	Age	Outcome
148	33.6	0.62	50	1
85	26.6	0.35	31	0
183	23.3	0.67	32	1
89	28.1	0.16	21	0
137	43.1	2.28	33	1

C. Multilayer Perceptron Classifier

A multilayer perceptron (MLP) is a class of feed-forward artificial neural network. We use this algorithm because MLPs are used in research for their ability to solve problems sarcastically, which often allows approximate solutions for extremely complex issues like fitness approximation.

Learning occurs in the perception by changing connection weights after each portion of data is processed, based on the quantity of error in the output associated with the exacted result. This is an example of supervised learning and is carried out through back-propagation, a generalization of the ”least mean squares algorithm” in the linear perception. We represent the error in output node j in the nth data point (training example) by,

$$e_j(n) = Y_j(n) - a(n) \tag{1}$$

Where Y is the target value, and a is the value produced by the perception. The node weights are adjusted based on corrections that minimize the error in the entire output, given by,

$$\epsilon(n) = \frac{1}{2}\sum_j e_j^2 \tag{2}$$

There are many hyper-parameters for MLP classifier such as alpha, hidden-layer size, solver, learning-rate decay, etc. To find the best model, the different combination of these hyper-parameters are tried randomly and iteratively. Firstly the model gets lower accuracy due to the high bias problem since it gives test and training accuracy pretty much same. There is some solution for high bias which is given below:

- Build a bigger network
- Train for a longer period
- Search for different NN (Neural network) architecture

The appropriate choice is option one and three because the training set is minimal so training for a longer time is not a very effective process to remove high bias problem. Analyzing the Pima dataset, we found the values of these following parameters. We selected LBFGS as an optimizer (an optimizer in the family of quasi-Newton methods), alpha=1e-5, and hidden layer sizes = (15, 7, 7, 3). In the hidden layer, the first layer has 15 node (neurons), second layer has seven neurons, third layer has seven neurons, and fourth layer has three neurons.

We apply logistic regression on these features, and we observed the accuracy level which is shown in Table II.

TABLE II
M.L.P. CLASSIFIER ACCURACY

Algorithm	Accuracy
Training Set: M.L.P. Classifier	86.73%
Test Set: M.L.P. Classifier	85.15%

VI. Applying K-means with different machine learning model

A. K-means Algorithm

Cluster analysis aims at partitioning the observations into disparate clusters so that comments within the same group are more closely related to each other than those assigned to different clusters [14]. Fig. 4 shows the procedure of the K-means algorithm, and the methods for the K-means Cluster algorithm is given below:

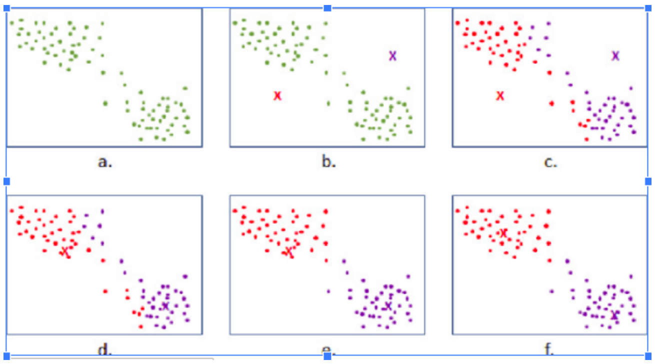

Fig. 4. Visualizing the k-means algorithm for Pima Indian data-set.

- Show all objects (step a). Select K from provided N as the number of initial cluster center (step b). In Fig. 4b, the value of K is 2.
- Calculate the distance between each object and cluster center.[6] (step c).
- Recalculate every cluster center to verify whether they are changed.
- Circulate step 2 and step 3 until the new cluster center is the same as the original one, i.e., convergence and end of the algorithm (step e and f).

Fig. 5 shows the output of the K-means clustering algorithm applied on Pima Indian Data-set.

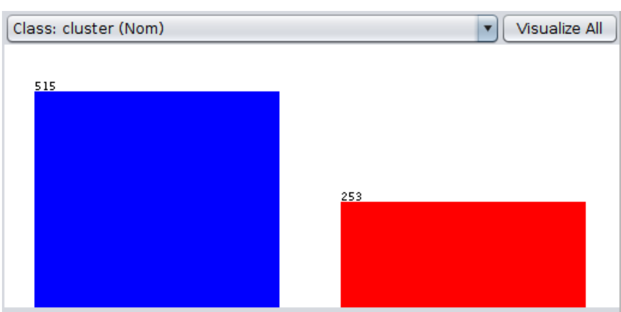

Fig. 5. k-means clustering result in Pima Indian data-set.(Blue for diabetes negative and red for diabetes positive)

K-means algorithm result can be used as an input feature which gives a good advantage in accuracy.

B. Feature Selection

The Greedy stepwise search is a feature selection algorithm (Discuss in Section V(B)) that can be used for selecting the useful feature. The greedy stepwise search algorithm takes those set of feature that gives minimum error rate. Here are the selected features.

- Pregnancies
- Glucose
- B.M.I.
- Age
- Diabetes Pedigree Function
- Cluster(Output of k-means algorithm)

C. Classifier

We apply many different kinds of the classifier to examine which method works well. Different classifier like the decision tree, MLP, and Logistic regression has a different approach to evaluate the model. Table III contain the result of the different classifier:

TABLE III
The result of the different classifier

Algorithm	Accuracy
Logistic Regression	77.08%
M.L.P. Classifier	75.39%
Random Forest	75.00%

Logistic regression is most suitable for this kind of problem because its cost function is targeted to make (zero, one) classifier. For this reason, this works well compared to M.L.P. and Random Forest.

VII. Performance Evaluation

The difference between two proposed methods is that one is using data recovery technique to eliminate noise and another one is using K-means based noise reduction technique. The first method showed improved accuracy for the given dataset. Here, we observed that the data recovery with the mean value more stable. But the second method which used K-means with neural network did not improve much as the data-set is noisy due to the different kinds of missing value. And therefore applying k-means for clustering the feature does not affect enhancing the efficiency of the classifier.

VIII. Discussion

A. M.L.P. classifier work

Multilayer Perceptron Classifier is a deep neural network classifier. It cannot be determined at the beginning which hyper-parameter like learning rate, batch size, optimizer; hidden layer size works best for the model. Only after analyzing the errors like high bias or high variance, the model can be built by an iterative process that gives a low error rate. In the beginning, M.L.P. classifier starts with hidden layer size=(16,8,2). The model then gives the same training set error and test set error. We found that the model is affected by high bias problem. So it is required to make the model big. After some iterative process and changing different hyper-parameters, an accuracy of 85% is achieved.

B. Comparison with others experiments

A comparison is made with work done by other researchers using the same data set in Table 5 where the table shows the result of the different work done by other researchers.

TABLE IV
COMPARISON WITH OTHERS EXPERIMENTS

Method	Accuracy	Reference
GRNN	80.21%	Kamer Kayaer [16]
Naive Bayes	78.69%	Dilip Kumar Choubey [17]
FMM with neural network	78.39%	Manjeevan seera [18]
J48graft	83.83%	Hayashi [19]
Hybrid model	84.5%	Humar Kahramanli [20]
MLP	81.9%	Aliza Ahmad [21]
ELM	75.72%	Rojalina Priyadarshini [22]
Artificial bee colony	84.21%	Beloufa [23]
Swarm intelligence	82.03%	Christopher [24]
fuzzy rule	79.37%	Lekkas [25]
K-means	77%	**Our approach**
MLP with Feature Selection	85.153%	**Our study**

IX. CONCLUSIONS

This paper presents an analysis of two kinds of the prediction model for diabetes mellitus and making the model adapt to different data-sets. Using these two methods, we can train a model that can predict whether or not someone has diabetes at a very early stage with the help of some features like Glucose level, BMI, Age.

The method incorporating k-means based noise reduction technique is easy to implement, but provides lower efficiency and requires more computational power. Whereas, the process that includes the data-recovery with neural network requires less computation, but offers the higher efficiency of 85%.So, the method that is capable of data-recovery with neural network is more acceptable.

For future work, it is essential to bring in the hospitals real and latest patients data for continuous training and optimization of our proposed model. The quantity of the data-set should be large enough to train appropriately and predict with higher efficiency [26], [27].

It is more useful and efficient for people to obtain an application about health management of DM on their mobile devices [28], [29]. An application can be developed that will provide rational and reasonable health advice to the high-risk group. Diabetes patients can be convinced to use this application to test their blood glucose level, blood pressure, and heart rate.

ACKNOWLEDGEMENT

We would like to show our gratitude to the Bangladeshi Engineers and Scientists in the USA specially Raihan Masud, S M Iftekharul Alam and Farzana Khalid for mentoring this research project and reviewing our paper as part of their extensive support from Ankur International, Portland, USA(https://ankurintl.org/project/about-aus-scholarships/). Also thanks to Manish Sah for his excellent suggestions.

REFERENCES

[1] International diabetes federation (idf) diabetes atlas (seventh ed.).

[2] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International journal of medical informatics*, vol. 77, no. 2, pp. 81–97, 2008.

[3] M. K. Hasan, M. Haque, N. Sakib, R. Love, and S. I. Ahamed, "Smartphone-based human hemoglobin level measurement analyzing pixel intensity of a fingertip video on different color spaces," *Smart Health*, vol. 5, pp. 26–39, 2018.

[4] M. K. Hasan, N. Sakib, J. Field, R. R. Love, and S. I. Ahamed, "A novel technique of noninvasive hemoglobin level measurement using hsv value of fingertip image," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, 2017, pp. 221–229.

[5] M. K. Hasan, N. Sakib, R. R. Love, and S. I. Ahamed, "Rgb pixel analysis of fingertip video image captured from sickle cell patient with low and high level of hemoglobin," in *Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017 IEEE 8th Annual*. IEEE, 2017, pp. 499–505.

[6] G. M. Ahsan, M. O. Gani, M. K. Hasan, S. I. Ahamed, W. Chu, M. Adibuzzaman, and J. Field, "A novel real-time non-invasive hemoglobin level detection using video images from smartphone camera," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, 2017, pp. 967–972.

[7] M. K. Hasan, M. Haque, R. Adib, J. Tumpa, A. Begum, R. Love, Y. L. Kim, and S. I. Ahamed, "Smarthelp: Smartphone-based hemoglobin level prediction using an artificial neural network," in *AMIA Annual Symposium Proceedings*, vol. 2018. American Medical Informatics Association, 2018, p. 535.

[8] F. Kawsar, M. K. Hasan, R. Love, and S. I. Ahamed, "A novel activity detection system using plantar pressure sensors and smartphone," in *Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual*, vol. 1. IEEE, 2015, pp. 44–49.

[9] F. Kawsar, M. K. Hasan, T. Roushan, S. I. Ahamed, W. C. Chu, and R. Love, "Activity detection using time-delay embedding in multi-modal sensor system," in *International Conference on Smart Homes and Health Telematics*. Springer, 2016, pp. 489–499.

[10] M. K. Hasan, G. M. T. Ahsan, S. I. Ahamed, R. Love, and R. Salim, "Pain level detection from facial image captured by smartphone," *Journal of Information Processing*, vol. 24, no. 4, pp. 598–608, 2016.

[11] K. H. Md, M. T. A. Golam, I. A. Sheikh, L. Rechard, and S. Reza, "Pain level detection from facial image captured by smartphone," *ł*, vol. 57, no. 7, 2016.

[12] M. Gittens, R. King, C. Gittens, and A. Als, "Post-diagnosis management of diabetes through a mobile health consultation application," in *e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on*. IEEE, 2014, pp. 152–157.

[13] D. K. Choubey and S. Paul, "Ga_mlp nn: a hybrid intelligent system for diabetes disease diagnosis," *International Journal of Intelligent Systems and Applications*, vol. 8, no. 1, p. 49, 2016.

[14] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2018.

[15] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, 2001, pp. 577–584.

[16] K. Kayaer and T. Yıldırım, "Medical diagnosis on pima indian diabetes using general regression neural networks," in *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, 2003, pp. 181–184.

[17] D. K. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," in *Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016)*, 2017, pp. 451–455.

[18] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2239–2249, 2014.

[19] Y. Hayashi and S. Yukita, "Rule extraction using recursive-rule extraction algorithm with j48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the pima indian dataset," *Informatics in Medicine Unlocked*, vol. 2, pp. 92–104, 2016.

[20] H. Kahramanli and N. Allahverdi, “Design of a hybrid system for the diabetes and heart diseases,” *Expert systems with applications*, vol. 35, no. 1-2, pp. 82–89, 2008.

[21] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, “Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus,” in *Digital Information Processing and Communications*. Springer, 2011, pp. 537–545.

[22] R. Priyadarshini, N. Dash, and R. Mishra, “A novel approach to predict diabetes mellitus using modified extreme learning machine,” in *Electronics and Communication Systems (ICECS), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[23] F. Beloufa and M. A. Chikh, “Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm,” *Computer methods and programs in biomedicine*, vol. 112, no. 1, pp. 92–103, 2013.

[24] J. J. Christopher, H. K. Nehemiah, and A. Kannan, “A swarm optimization approach for clinical knowledge mining,” *Computer methods and programs in biomedicine*, vol. 121, no. 3, pp. 137–148, 2015.

[25] S. Lekkas and L. Mikhailov, “Evolving fuzzy medical diagnosis of pima indians diabetes and of dermatological diseases,” *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 117–126, 2010.

[26] H. Li, Q. Zhang, and K. Lu, “Integrating mobile sensing and social network for personalized health-care application,” in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 2015, pp. 527–534.

[27] Y. Luo, C. Ling, J. Schuurman, and R. Petrella, “Glucoguide: An intelligent type-2 diabetes solution using data mining and mobile computing,” in *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*. IEEE, 2014, pp. 748–752.

[28] R. Schnall, M. Rojas, S. Bakken, W. Brown III, A. Carballo-Dieguez, M. Carry, D. Gelaude, J. P. Mosley, and J. Travers, “A user-centered model for designing consumer mobile health (mhealth) applications (apps),” *Journal of biomedical informatics*, vol. 60, pp. 243–251, 2016.

[29] M. A. Basar, H. N. Alvi, G. N. Bokul, M. S. Khan, F. Anowar, M. N. Huda, and K. A. Al Mamun, “A review on diabetes patient lifestyle management using mobile application,” in *proceeding of the 18th International Conference on Computer and Information Technology*, 2015, pp. 379–385.