Sequential Minimal Optimization for Predicting Diabetes at its Early Stage

Apeksha Khanwalkar
Department of Computer science engineering Institute of Technology and Management
Gwalior, India
Apeksha.Khanwalkar@gmail.com
Dr.Rishi Soni
Department of Computer science engineering Institute Of Technology and Management
Gwalior, India
Rishi.soni@itmgoi.in

***Abstract*—Diabetes is a chronic disease that pays for a large proportion of the nation's healthcare expenses when people with diabetes want medical care continuously. It is essential to recognize high-risk populations, performance & modifications to prevent and delay the onset of Type 2 diabetes. Most diabetics do not know much about their health or the factors of risk before diagnosis. In this work, the WEKA tool applied to the Pima Indians Diabetes dataset that is used to compare our outcomes with the existing outcomes. In the previous work, three classification algorithms were used among which Naïve Bayes stand to be the most efficient one. But, there were some limitations in the Naïve Bayes algorithm hence Sequential Minimal Optimization algorithm (SMO) is used. A comprehensive review of literature has been introduced which highlights the research operations using various data mining algorithms and tools as well. The comparison graph represents that the proposed model achieved a higher accuracy prediction rate in comparison to previous work.**

***Keywords*—Data Mining, Diabetes Data, Chronic Disease, WEKA, SMO**

I. INTRODUCTION

Over the past two decades, essential mechanisms linked to insulin resistance, active diabetes, and associated cardiometabolic diseases have been clarified considerably. Nevertheless, more work is required to further understand the insulin resistance's pathophysiologic profile in cases where diabetes is co-morbid with other chronic diseases.

The extraction of useful information from this large dataset takes place in data mining, which has played a significant role in the field of safety. Procedures for data mining for diabetic academics should be welcomed, as secret knowledge can be uncovered from an immense amount of diabetes-related data[3]. Data mining is the way that a vast number of databases have valuable knowledge and trends. The data set is broad, complex and difficult to forecast based on statistics, especially on health care. Several data mining approaches are proposed in the sense of the clinical field. Statistical analysis, estimation, and classification are the most common and well-used data mining techniques. With an enormous amount of medical data gathered from various sources, it is increasingly more relevant for analytics, analysis, and decision-making to be established with more powerful apps. This method applies to indirect retrieval of data from unknown/hidden or unique data from the database.

Diabetes Mellitus has a prevalence of type 2 diabetes over the past decade which tends to rise due to the increase of the aging population, urbanization as well as a highly obese life [1]. The potential for more health complications like kidney failure, heart disease, blindness and neurological damage, ultimately fatal, is the effect of diabetes on human development. Moreover, the result of the disease on one population is the lack of a stable working workforce and expanded health spending, rendering diabetes one in several nations, including Thailand, another of public health issues. [2]. In Thailand, health expenses incurred as a result of diabetes amounted to 11% of the total health spending in 2010[3]. Diabetes public health strategies are intended to treat the condition, and thus mangrove, improved treatment, as therapy trials have demonstrated the dietary or behavioral improvements in people at high risk will avoid diabetes [4],[5]. need to distinguish high-risk people from the rest of the population, i.e. screening, specifically and cheaply is, therefore, a significant issue and this paper concentrates.

Current diabetes diagnosis methods rely on laboratory tests including fasting blood glucose and resistance to oral glucose. Nevertheless, both invasive and time-consuming procedures are unsuitable for the population to be tested. Such laboratory experiments, in particular, are not useful after the illness starts, which is too late to be considered in advance for successful research.

In previous years, the diabetes studies focused on surveys using methods for risk scoring to measure the probability of diabetes without including laboratory tests. Danger scoring has shown good prevision, high sensitivity, and is fairly easy to completely measure by hand [6].

Diabetes is a long-lasting illness affecting humans all over the world. This occurs because the body is not able to produce enough insulin. To regulate the level of glucose, insulin is released by the pancreas, one of the most essential hormones in the body. Despite insulin injections, balanced diets or daily exercise, diabetes can not be regulated. Blindness, pain from blood, heart problems, kidney failure, etc. are causes of diabetes. Four forms of diabetes are mostly found:

Type 1 Diabetes: This develops when insulin can not be released by the pancreas. Insulin is a hormone released by the pancreas. At any age, type 1 diabetes can occur. It happens most often among children and young people [7]

Type 2 Diabetes: This occurs because insulin is not enough to satisfy the needs of blood. Obesity raises the risk of Type 2 diabetes, primarily arising at 40 years of age, owing to family origin, weight. [3]

Gestational Diabetes: It is the third form that generally done because of the level of high blood glucose quantity present in the body in pregnant women. [4]

Pregestational Diabetes: Previously become pregnant insulin-dependent diabetes arises, this is pregestational diabetes. [6] In the data mining method the diabetes forecast shows an important part. For the forecast of diabetes, there are various data mining techniques are useful.

Data mining can use in the research of diabetes and eventually expand the quality of health care of diabetes victims.

II. LITERATURE REVIEW

The association between the composition of free fatty acids or metabolic parameters was studied but serum linoleic acid was shown to have a negative correlation to intestinal fat or insulin resistance. [8].

The effects of overexpressing gamma-glutamyl transferase on insulin sensitivity were examined and the short-term overexpression of liver-specific gamma-glutamyl transferase found insulin sensitivity. The liver illness usually happens in diabetes; other organ difficulties with cardiomyopathy, neuropathy, and nephropathy are recognized extremely is presented in [9].

The perception of the properties of renal denervation on insulin sensitivity in nondiabetic victims with behavior for resistant hypertension is presented in [10].

To reduce the gene profiles of neurotrophin-mitogen-activated protein kinase (MAPK) signifying in peripheral neuropathy patients with the diabetic is presented in [11].

It is examined how fast restoration of energy metabolism affected by normoglycemia during exercise in a nonobese patient with type 2 diabetes. The insulin-induced normoglycemia increased blood glucose during successive exercise without altering largely substrate usage is presented in [12].

The benefits of the game to enhance insulin sensitivity and glucose metabolism and suggested in [13] a therapeutic role of exercise-induced myokine & irisvin in the cooperative effects of exercise on glucose metabolism. During the exercise iris in of the protein produced in the muscles.

The authors have driven in [14] that Roux-en-Y gastric bypass operation is an adequate action to decrease the risk of cardiovascular amid diabetic victims with comorbid of overweight.

Apart from surgery, in this particular issue, another emerging therapeutic route has been introduced in [15], there is a review article where the authors highlighted the execution of the nuclear-factor-E2-related-factor antioxidant response element (ARE) method as an objective for curing, diagnosis, and medication of type 2diabetes.

The role of ML methods for ADOOST or Bagging ensemble [16] in the J48 decision tree as a framework for the classification of diabetes mellitus or diabetics based on diabetes risk factors was addressed. Reports after the trial show that the Adaboost ML ensemble strategy beats comparatively well bagging and a decision-making tree J48.

diabetes prediction system was developed, the primary goal of which is to prevent diabetes at a specific age of a candidate. The program proposed is constructed by adding a decision tree based on the principle of machine learning. The findings were encouraging, as the program is working well in forecasting diabetes events at a given age, and a decision tree is used to provide better accuracy in [17].

III. METHODOLOGY

For prognosticating the likelihood of diabetes in patients with maximum accuracy, it is necessary to find the efficient technique of data mining which helps to assess the accuracy for a given set of data. This research work focuses on pregnant women suffering from diabetes; therefore we have applied a classification algorithm on the Pima Indian diabetes dataset that provides better prediction results of disease.

In the previous work, three classification algorithms were applied on the Pima dataset among which Naïve Bayes outperformed the best. But this algorithm has some limitations which are overcome by the new research work. In the proposed research, we have utilized new training algo that train support vector machines, i.e. Sequential Minimal Optimization or SMO.

NAIVE BAYES

NB is a Bayesian supervised classifier that adopts that all attributes are self-regulating of each other, which is referred to in class. Therefore, for the figure of our data distribution Naive Bayes classifier generates an extremely strong assumption, i.e. any two structures are independent of the given output class. The result can be (potentially) very bad, because of this. Due to data scarcity, another problem will occur. For any probable value of any feature, you need to constantly anticipate the possibility of value by the frequentist approach. This can be the result with the probabilities of to 0 or 1, result in numerical instabilities and bad results.

SEQUENTIAL MINIMAL OPTIMIZATION

- Train the dataset by using the support vector machine is vital to establish a very huge quadratic programming (QP) optimization difficulty. This big QP problem breaks into a series of small potential QP problems by the SMO. These small QP problems are systematically determined, which avoids a time-consuming mathematical QP optimization as an internal loop. Managing the large training data 0f SMO allows, in the training set size needs the size of storage for SMO is linear. The advantages of SMO are that it yields the detail that the production of 2 Lagrange multipliers could be done analytically. Since SVMs fittings are used on security, which does not necessarily depend on the number of structures, SMO can handle large structure spaces.

ALGORITHM: SMO

Step 1: Collect data set from -d uciml/Pima-Indians-diabetes-database

Step 2: Convert CSV data format to arff format

Step 3: Start

Step 4: Load dataset

Step 5: Convert numeric dataset into nominal form

Step 6: Now apply the SMO algorithm

Step 7: Run the algorithm

Step 8:Obtain the predicted results

Step 9: End the process

This algorithm can also be visualized in the form of a flow diagram that has been designed below highlighting every step involved in the research process.

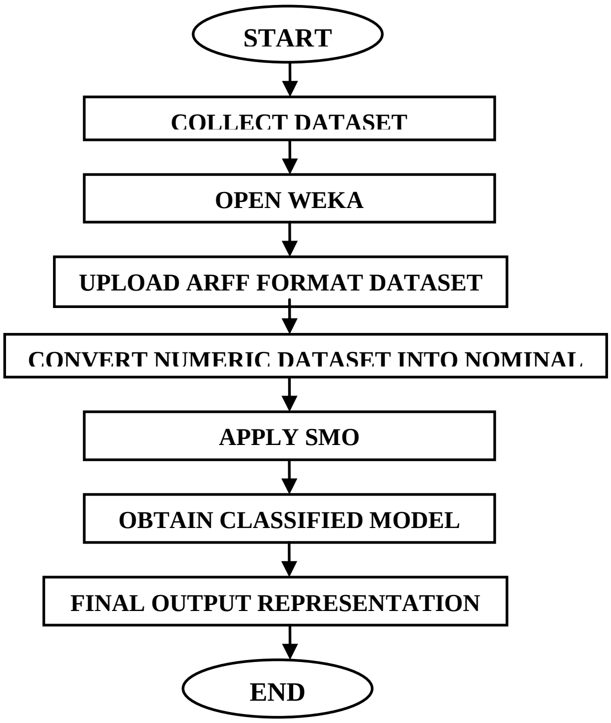

START
COLLECT DATASET
OPEN WEKA
UPLOAD ARFF FORMAT DATASET
CONVERT NUMERIC DATASET INTO NOMINAL
APPLY SMO
OBTAIN CLASSIFIED MODEL
FINAL OUTPUT REPRESENTATION
END

Fig. 1 Dataflow diagram

The above flow diagram in fig.1 represents various steps involved in the working process of SMO. The steps of this algorithm are discussed in the flow chart designed above where every step has been highlighted.

IV. EXPERIMENTAL RESULTS AND ILLUSTRATIONS

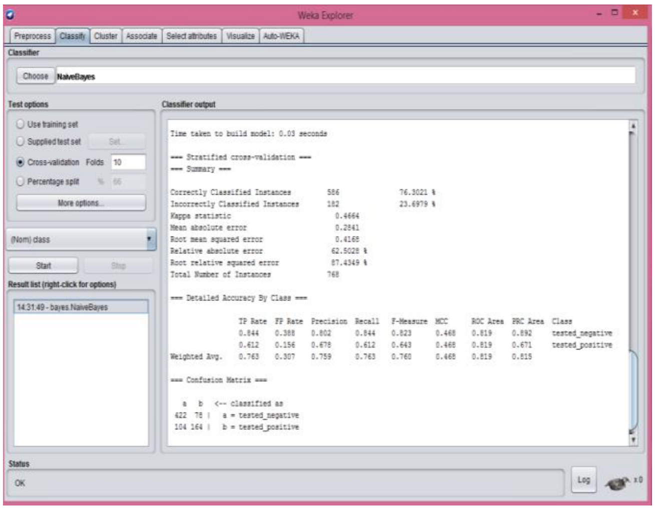
Analyzing the dataset for any resulting analysis is the main objective of any experimental setup. The projected methodology is occupied from the UCI Repository which is a valuedat Diabetes Dataset (PIDD). This sample consists of 768 female patients in medical detail. the dataset contains 8 numerically assessed attributes where the value of one diabetes' 0' class as well as another diabetes class' 1' are viewed as valid. Table-I representing Information Attributes is known as a data set definition.

Table I. Pima Indians diabetes dataset

S. no.	Attribute	Description
1	Pregnancies	Number(denote how many times the pregnancy occur)
2	Glucose	Plasma glucose concentration 2 hours in an oral glucose tolerance test
3	Blood Pressure	In mm Hg(Diastolic blood pressure)
4	skin thickness	In mm (Triceps skinfold thickness)
5	Insulin	mu U/ml(2-Hour serum insulin)
6	BMI	weight in kg/(height in m)^2(Body mass index)
7	Diabetes Pedigree Function	Diabetes pedigree function
8	Age	In years
9	Outcome	The class variable has 2 values (0 or 1) 268 of 768 are 1, the others are 0

Here for analyzing our dataset, we used the WEKA 3.9.0 tool. WEKA method developed in New Zealand by the University of Waikato is used to build algos from data mining. WEKA is an innovative machine learning (ML) research-creation facility and its uses in real-world data mining challenges.

In current years, the use of data mining algorithms has increased in medical forecasting analysis because of the earnest research in related areas. By this WEKA toolkit & the Pima Indian Diabetes dataset, some scholars have achieved numerous considerable results. However, there is scope for improving accuracy.
Fig. 2 shows the final predicted outcomes of NB algo after training the PIDD dataset. the result shows 76.3021% accuracy and the time is taken for the processing of this algorithm is 0.03 seconds.



Classifier: Choose NaiveBayes

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

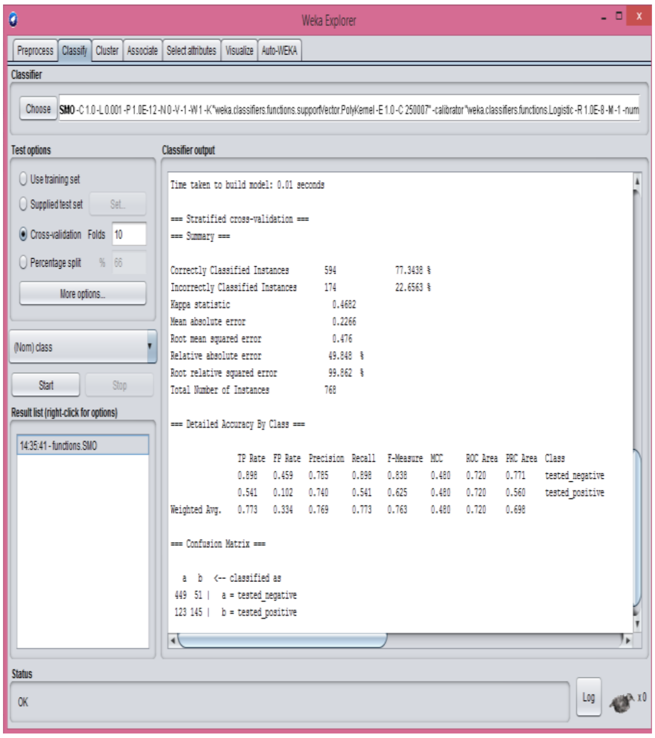
Result list (right-click for options): 14:31:49 - bayes.NaiveBayes

Status: OK

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.844    0.388    0.802      0.844    0.823      0.468    0.819     0.892     tested_negative
                 0.612    0.156    0.678      0.612    0.643      0.468    0.819     0.671     tested_positive
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468    0.819     0.815

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

Fig. 2 Result and Visualization of Naïve Bayes

Fig. 3 shows the accuracy of Sequential minimal optimization over the existing research models i.e., Naïve Bayes. A comparison has been made which shows that the proposed algorithm has better accuracy than that of the existing algorithm.



Classifier: Choose SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) class

Result list (right-click for options): 14:35:41 - functions.SMO

Status: OK

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         594               77.3438 %
Incorrectly Classified Instances       174               22.6563 %
Kappa statistic                          0.4682
Mean absolute error                      0.2266
Root mean squared error                  0.476
Relative absolute error                 49.848  %
Root relative squared error             99.862  %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.898    0.459    0.785      0.898    0.838      0.480    0.720     0.771     tested_negative
                 0.541    0.102    0.740      0.541    0.625      0.480    0.720     0.560     tested_positive
Weighted Avg.    0.773    0.334    0.769      0.773    0.763      0.480    0.720     0.698

=== Confusion Matrix ===

   a   b   <-- classified as
 449  51 |   a = tested_negative
 123 145 |   b = tested_positive
```

Fig. 3 Result and visualization of SMO

Table II. Comparison of Training and Simulation Error

Evaluation Criteria	**Classifiers**	
	Naïve Bayes	SMO
Kappa Statistics (KS)	0.4664	0.4682
Mean Absolute Error (ABE)	0.2841	0.2266
Root mean squared error (RMSE)	0.4168	0.476

Relative absolute error % (RAE)	62.5028%	49.848%
Root relative squared error % (RRSE)	87.4349%	99.862%

Table II shows the comparison of simulation errors and the training provided to both the research work using the same data where SMO proves to be the better one.

Table III. Comparing the performance of the students

Algorithm	Correctly classified instances %	Incorrectly classified instances %	Execution Time (Seconds)
Naïve Bayes	76.3021%	23.6979%	0.03
SMO	77.3438%	22.6563%	0.01

Table III shows the values of correctly and incorrectly classified instances along with the time duration it took. This comparison clearly shows that the SMO algorithm can perform better than Naïve Bayes with a short duration of time.

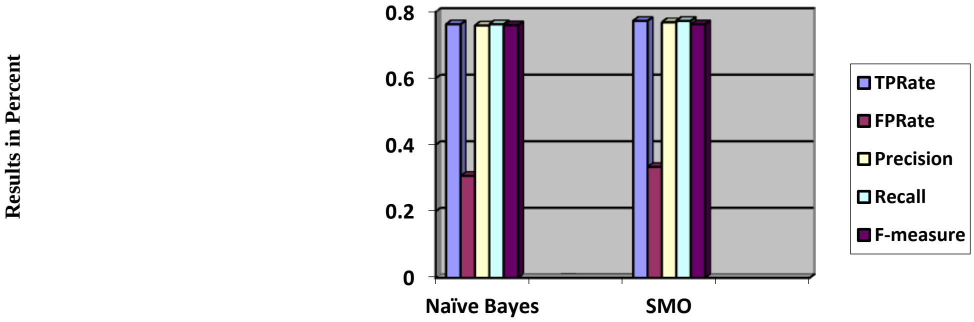


Fig. 4 Comparing the accuracy of classifiers in Base and Propose work

Fig. 4 shows the classifier accuracy of the base and propose work with SMO method having higher values than that of the Naïve Bayes.

V. CONCLUSION

Early prediction of any disease is the main concern. Here, we have talked about diabetes which is a real-world problem. However, various researches have been made in this field but still, this is a challenging issue. In this paper, the SMO classifier is used to predict those patients who are suffering from diabetes in their early stages. In the existing research, the Naïve Bayes classification algorithm is used which SMO has replaced in the new research. SMO has proven better than NB in terms of accuracy & time taken to perform on a dataset. The proposed approach achieved 77.34% accuracy while in an existing approach it was 76.30%. Also, the time taken by SMO is less than Naïve Bayes.

In the future, we can use some other techniques of data mining to gain more powerful results. The early prediction of the disease will predict via another simulation tool.

References

[1] J. Shaw, R. Sicree, and P. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," Diabetes Research and Clinical Practice, vol. 87, no. 1, pp. 4–14, Jan. 2010.

[2] Bureau of Non-Communicable Disease, Ministry of Public Health, Thailand, NCD Annual Report 2013, 1st ed., 2013.

[3] P. Zhang, X. Zhang, J. Brown, D. Vistisen, R. Sicree, J. Shaw, and G. Nichols, "Global healthcare expenditure on diabetes for 2010 and 2030," Diabetes Research and Clinical Practice, vol. 87, no. 3, pp. 293–301, Mar. 2010.

[4] XR Pan et al., "Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance: the Da Qing IGT and Diabetes Study," Diabetes Care, vol. 20, no. 4, pp. 537–544, Apr. 1997.

[5] Tuomilehto, Jaakko et al., "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance," New England Journal of Medicine, vol. 344, no. 18, pp. 1343–1350, 2001.

[6] J. Lindstrom and J. Tuomilehto, "The Diabetes Risk Score A practical tool to predict type 2 diabetes risk," Diabetes Care, vol. 26, no. 3, pp. 725–731, Mar. 2003.

[7] Diabetes Association of Thailand, The Endocrine Society of Thailand, Department of Medical Services, and National Health Security Office, Clinical Practice Guideline for Diabetes 2014, 1st ed. AroonKarnPim, 2014.

[8] M. Mishra and J. F. Ndisang, "A critical and comprehensive insight on heme oxygenase and related products including carbon monoxide, bilirubin, biliverdin and ferritin in type-1 and type-2 diabetes," Current Pharmaceutical Design, vol. 20, no. 9, pp. 1370–1391, 2014.

[9] J. F. Ndisang and A. Jadhav, "Hemin therapy improve kidney function in male streptozotocin-induced diabetic rats: role of the heme oxygenase/atrial natriuretic peptide/adiponectin axis," Endocrinology, vol. 155, no. 1, pp. 215–229, 2014.

[10] M. C. Petersen, D. F. Vatner, and G. I. Shulman, "Regulation of hepatic glucose metabolism in health and disease," Nature Reviews Endocrinology, vol. 13, no. 10, pp. 572–587, 2017.

[11] J. F. Ndisang, A. Jadhav, and M. Mishra, "The heme oxygenase system suppresses perirenal visceral adiposity, abates renal inflammation and ameliorates diabetic nephropathy in Zucker diabetic fatty rats," PLoS One, vol. 9, no. 1, article e87936, 2014.

[12] P. Hossain, B. Kawar, and M. El Nahas, "Obesity and diabetes in the developing world—a growing challenge," The New England Journal of Medicine, vol. 356, no. 3, pp. 213–215, 2007.

[13] M. Vladu, D. Clenciu, I. C. Efrem et al., "Insulin resistance and chronic kidney disease in patients with type 1 diabetes mellitus," Journal of Nutrition and Metabolism, vol. 5, 2017.

[14] J. F. Ndisang and A. Jadhav, "Heme arginate therapy enhanced adiponectin and atrial natriuretic peptide, but abated endothelin-1 with attenuation of kidney histopathological lesions in mineralocorticoid-induced hypertension," The Journal of Pharmacology and Experimental Therapeutics, vol. 334, no. 1, pp. 87–98, 2010.

[15] L. Duvnjak and M. Duvnjak, "The metabolic syndrome: an ongoing story," Journal of Physiology and Pharmacology, vol. 60, Supplement 7, pp. 19–24, 2009.

[16] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," Procedia Computer Science, vol. 82, pp. 115–121, 2016. DOI:10.1016/j.procs.2016.04.016.

[17] Orabi, K.M., Kamal, Y.M., Rabah, T.M., "Early Predictive System for Diabetes Mellitus Disease," in Industrial Conference on Data Mining, Springer, pp. 420–427, 2016.