# Gestational Diabetes Diagnosis with MSVM, MJ48 Classifier Models

S. Saradha, P. Sujatha

***Abstract--- This paper focuses on designing an automated system for diagnosing gestational diabetes. Classification is one of the common predictive data mining tasks. It arranges the information and assembles a model to deliver the new grouped information. 'Gestational diabetes mellitus' (GDM) is a form of diabetes that occurs during pregnancy due to hormonal changes. Pregnant Women with GDM are at highest threat of future diabetes, especially type-2 diabetes. To diagnose the GDM, the two classifier models are proposed such as .Modified Support Vector Machine (MSVM) and Modified J48 (MJ48). Based on the performance analysis, the classifier model MJ48 provides more accuracy and less error rate than MSVM proposed classifier model.***

***Keywords--- Data Mining, Classifiers, GDM, OGTT, MSVM, MJ48, Accuracy.***

## I. INTRODUCTION

1. **Data Mining:** Knowledge plays a significant role in every stage of human life development. To gain this knowledge, one has to analyze the available unlimited data in different formats present in the database. It is possible to analyze this data and reveal the hidden information using the concept of 'data mining.' The term 'data mining' means a method or process that can extract interesting knowledge from a given large data.
2. **Medical Data Mining:** Clinical databases are the most extensive data storing fields across the globe. Identifying the patterns and relationships from these extensive data can help medical caretakers with new medical knowledge. Unfortunately, there are very few methods and systems that can perform this task. This study focuses on applying data mining techniques in the clinical database and identifies some interesting facts from it.
3. **Classification:** Classification means to divide or group data into classes. A classification technique is to identify each dataset and put them in the right target class. There is a very minimal difference between clustering and classification. Classification comes under 'supervised learning' and clustering comes under 'unsupervised learning method.'
4. **Diabetes Mellitus:** Diabetes mellitus (DM) is a persistent disease and an essential global communal health confront. It occurs when a body is not in a position to respond or consequence appropriate to insulin, which is wished to hold the charge of glucose. Diabetes can be inhibited with the aid of insulin injections, a healthy weight loss plan, and regular exercising but there is no total therapy obtainable. Diabetes prompts various diverse sicknesses, for example, visual impairment, pulse, coronary illness, kidney ailment, and nerve harm. Extreme diabetes may likewise prompt advance exorbitant risk reason for passing on. Type 1 DM, Type2 DM and 'Gestational Diabetes Mellitus' (GDM) are the three primary types of Diabetes Mellitus'.
5. **Gestational Diabetes Mellitus (GDM):** 'Gestational diabetes' affects the female only in the time of pregnancy. Gestational diabetes happens in around 5% all things considered. If it is untreated, it may cause many health problems for mother and fetus-like miscarriage, preterm birth, fetal death, and congenital malformation. Age, previous unknown stillbirth, family history, and excess weight are some of the reasons for GDM. Various screening tests are done at the different weeks of pregnancy to diagnose diabetes. At the first prenatal routine fasting, glucose measurement screening test will be done to assess GDM. If the fasting glucose level is less than 5.1mmol/l, then it is treated as normal. Suppose the glucose level is greater than 7.0mmol/l, it is suspected as pregestational diabetes. For suspected GDM the glucose level should be in the range of 5.1-7.0 mmol/l. 'Oral Glucose Tolerance Test (OGTT)' should be done at the 24 weeks of gestation. If the GDM symptoms are present, the OGTT should be done again.

## II. LITERATURE SURVEY

S. Kavipriya et al., [10] discussed diabetes as a part of growing diabetes. Gestational diabetes mellitus becomes highly prevalent disorder these days among pregnant women. The GDM can be associated with maternal and prenatal outcomes.

The GDM treatment becomes full - team benefit becomes a short-term treatment.

Mani Butwall et al., [14] discussed Diabetes mellitus becomes an endless disease, forces excessively high social and financial expenses for a nation. The additional information was minimizing the standard rate. Further excessive and risky confusions required a variable administration such as the diabetes administration focus on the secure participation between the patient and health awareness experts. The data mining gave the diversity of methods, which investigates huge data to find the hidden knowledge.

Manuscript received September 16, 2019.

**Dr.S. Saradha,** Assistant Professor, Department of Computer Science, VISTAS. (e-mail: saradha.research@gmail.com)

**Dr.P. Sujatha,** Professor, Department of BCA & IT, VISTAS. (e-mail: suja.research@gmail.com)

D.SheilaFreeda et al., [15] analyzed with previous researches, the relationships for the different opportunity which may exist in increasing the diagnosis for effective treatment using data mining tools and techniques.

Vimalavinnarasi.et al., [12] explained diabetes as never-ending disease. The diseases which affect the major organs of the human body like heart, blood vessels, nerves, eyes, and kidneys, etc., The ‘World Health Organization’ (WHO) estimates such that nearly 200 million people around the world. Many of them have such diabetes. The number doubled by 2030. In recent research, India crosses almost 50 million diabetics, according to the statistics of the International Diabetes Federation. The identification for diabetes mellitus requires the medical practitioner, gives the diagnoses pattern consists of observable symptoms, which based on the test. The risk and costs differing based on the patient condition. The proposed work focused on the novel approach like a medical practitioner. The provision for some suggestions regulates the blood sugar level. The notification for risk factor relates the patient like heart attack, nerves problem. The affection of eye or kidney becomes the major problem. There may be two different modern approaches needed for the development of an automated model. The C5.0 algorithm uses the classification for the patient data and also focused for the fuzzy inference for analyzing data. The achievement of accurate results was examined through the analyzed data.

A.A. Ojugo [13] deals with the Diabetes Mellitus (silent killer or sugar disease) represented the metabolic syndrome characterized using the high glucose levels in the body with low insulin to break glucose. The body may be resistant to the effects of insulin. With the improvement in early diagnosis, data mining tools can get the better classification of the disease for endocrinologists. The present study represents the neural network model can train the hybrid fuzzy, genetic algorithm with the decision support system for diabetes classification.

## III. PROPOSED MODELS

For this research work, data sets of pregnant women are collected from ‘THEMBAVANI Hospital’ and ‘AASARAA Diagnostics’. With the received data set, preprocessing is done to eliminate zero and missing values. Discretize filter is used to replace these zero values with the proper value. The preprocessed data is uploaded in weka in CSV format. To diagnose GDM, three different classifier models such as MSVM and MJ48 are implemented and classify the given data set as GDM or Non- GDM. The performance analysis of three proposed models is carried out based on the parameters such as time taken to build the model, error rate, accuracy, etc. and the best classifier is identified. Also, the performance analyses of proposed methodologies are carried out with the existing methods based on the efficiency.

## IV. IMPLEMENTATION OF MSVM

MSVM proposed model is implemented by using the pre-processed data set. By using supervised Discretization technique, all the values of numeric attributes are converted to nominal. By modifying the values of parameters (such as ‘C' and ‘צ') of the existing SVM classifier, the proposed MSVM yields higher accuracy than the existing one. The following Figure explains the flow of the MSVM classifier model.

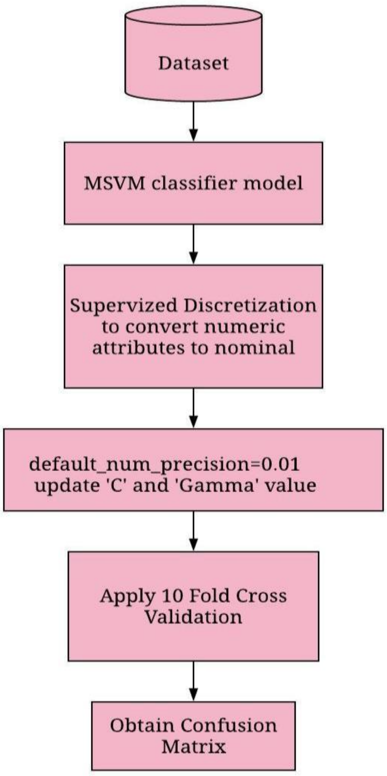
Dataset
MSVM classifier model
Supervized Discretization to convert numeric attributes to nominal
default_num_precision=0.01
update 'C' and 'Gamma' value
Apply 10 Fold Cross Validation
Obtain Confusion Matrix

**Figure 1: MSVM Classifier Model**

The confusion matrix of MSVM classifier model and its explanations is as follows:

**Table 1: Confusion Matrix of MSVM Algorithm**

| Total Number of Instances: | | 6629 | |
|---|---|---|---|
| **Sl.No.** | **Parameter** | **Detected** | |
| | | **Positive** | **Negative** |
| 1. | **Positive** | 910 (TP) | 585 (FN) |
| 2. | **Negative** | 309 (FP) | 4825 (TN) |

Total number of instances = 6629

Accuracy = (TP+TN) / (TP+TN+FP+FN) = (910+4825) / (910+4825+309+585) =86.5138%

Error rate = (FP+FN) / (TP+TN+FP+FN) = (309+585) / (910+4825+309+585) = 13.4862%

Precision = TP / (TP + FP) = 910 / (910+309) = 0.74

Recall = TP / (TP + FN) = 910 / (910+585) = 0.60.

**Table 2: Performance of MSVM Classifier Model**

| **Classifier** | **Time Taken to Build a Model** | **Accuracy Value (In %)** | **Error Rate** | **Precision** | **Recall** |
|---|---|---|---|---|---|
| **MSVM** | **0.06 seconds** | **86.51%** | **0.13** | **0.74** | **0.60** |

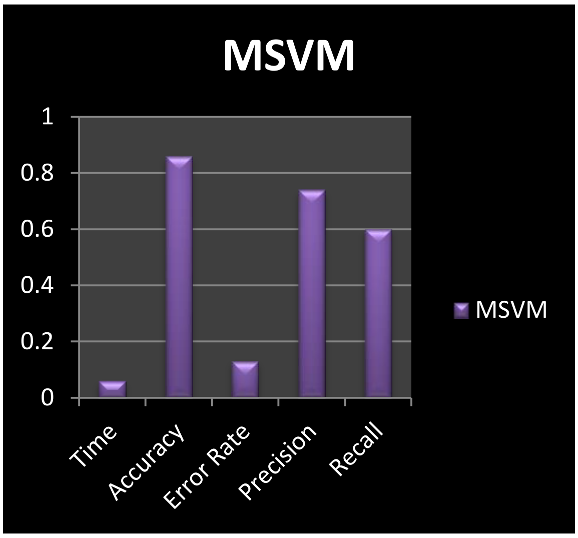

**Figure 2: Performance of MSVM Classifier Model**

From the confusion matrix of MSVM, it is observed that 6629 instances are considered for analysis out of which 5735 instances are correctly classified, and 894 cases are incorrectly classified. Table 2 and Figure 2 elaborately explain the performance measures of proposed MSVM classifier model.

## V. IMPLEMENTATION OF MJ48

MJ48 classifier model generates a decision tree for classifying the given dataset as GDM or Non-GDM. The algorithm of MJ48 is explained in Figure 3. For each attribute A, the parameters like confidence factor, the minimum number of objects, number of folds and seed values are assigned. Then the probability of a value 'A' will be found out, and a decision tree was formed. This process is repeated for other attributes also.

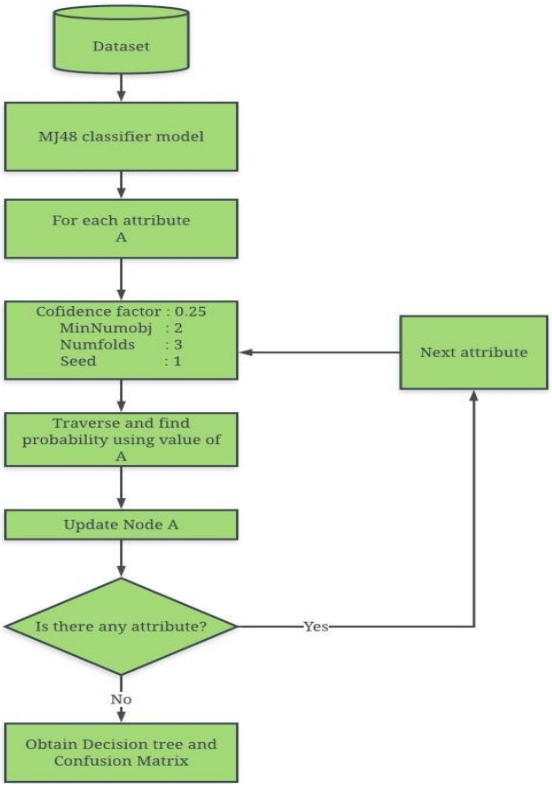

**Figure 3: MJ48 Classifier Model**

The proposed classifier MJ48model yields more accuracy than existing J48 by modifying some of the parameters which are explained as follows.

Pruning task is done for generalization of the tree. For improved pruning, in MJ48 the parameters such as Confidence Factor, MinNumobj, Numfolds and seed values are modified as 0.25, 2, 3 and one respectively. In existing J48 the costs of those attributes are 1.15, 2, 10 and one respectively. Below steps are performed in the proposed classification algorithm.

The 16-bit representation of the device MAC address is added in the Current Active Directory List. The MJ48 decision tree algorithm is used to examine the normalized information gain that is derived from identifying an attribute for splitting the data. To make a decision, the highest standardized information gain attribute is chosen. Then the algorithm moves to the smaller subsets. The splitting steps are terminated if all instances from a subgroup belong to the same class. Once this occurs, a leaf node is created denoting to select that class. In the case of MJ48 decision tree algorithm, a decision node is formed in the higher up of the tree with the help of the expected value of the class. The modified J48 decision tree is demonstrated in the following Figure. The first level holds the single header node, and it just acts as a pointer node for the children. The second level contains 2 sub trees named as 1 and 2.

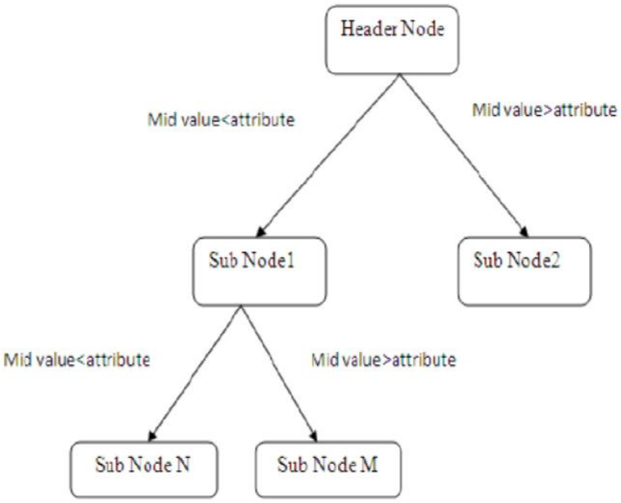

**Figure 4: Structure of MJ48 Decision Tree**

**Table 3: Confusion Matrix of MJ48 Algorithm**

| Total Number of Instances: | | 6629 | |
|---|---|---|---|
| | | Detected | |
| Sl.No. | Parameter | Positive | Negative |
| 1. | **Positive** | 1397 (TP) | 98 (FN) |
| 2. | **Negative** | 111 (FP) | 5023 (TN) |

Total number of instances = 6629

Accuracy = (TP+TN) / (TP+TN+FP+FN) = (1397+5023) / (1397+5023+111+98) =96.8472%

Error rate = (FP+FN) / (TP+TN+FP+FN) = (111+98) / (1397+5023+111+98) = 3.1528%

Precision = TP / (TP + FP) = 1397 / (1397+111) = 0.92

Recall = TP / (TP + FN) = 1397 / (1397+98) = 0.93

**Table 4: Performance of MJ48 Classifier Model**

| Classifier | Time Taken to Build a Model | Accuracy Value (in %) | Error Rate | Precision | Recall |
|---|---|---|---|---|---|
| MJ48 | 0 .12 seconds | 96.84% | 0.03 | 0.92 | 0.93 |

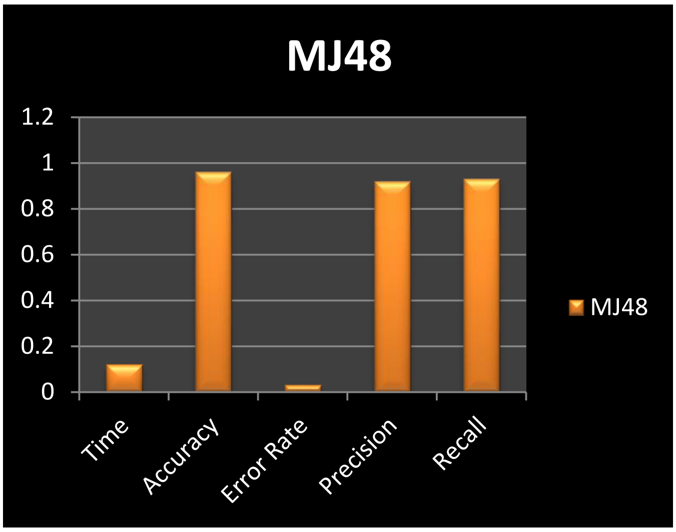

**Figure 5: Performance of MJ48 Classifier Model**

From the confusion matrix of MJ48, it is noticed that 6629 instances are considered for analysis out of which 6420 cases are correctly classified, and 209 instances are incorrectly classified. Table 4 and Figure 5 elaborately explains the performance measures of proposed MSVM classifier model.

## VI. RESULTS AND DISCUSSION

It is crucial to decide that one classification algorithm is superior to another. At times the classification algorithm that works well for a particular type of data may not work in the same behavior of another kind of data. Therefore these classification algorithms are evaluated based on accuracy. Accuracy is defined as the proportion of the total number of correctly predicted cases. The efficiency achieved by both model is high. MJ48 has better accuracy than MSVM. The database that is being used in this method is noise free and is large. So efficiency is said to be high. In general in case of machine learning approach data collection is more accurate. Figure 6 and Table 5 demonstrates the Time taken to build a model, Accuracy, error rate, precision, and recall for all the two proposed classifiers. These are the major parameters to evaluate the best classifier model.

**Table 5: Comparative Analysis of MSVM and MJ48 Classifier Model**

| Classifier | Time Taken to Build a Model | Accuracy Value (in %) | Error Rate | Precision | Recall |
|---|---|---|---|---|---|
| MSVM | 0.06 seconds | 86.51% | 0.13 | 0.74 | 0.60 |
| MJ48 | 0 .12 seconds | 96.84% | 0.03 | 0.92 | 0.93 |

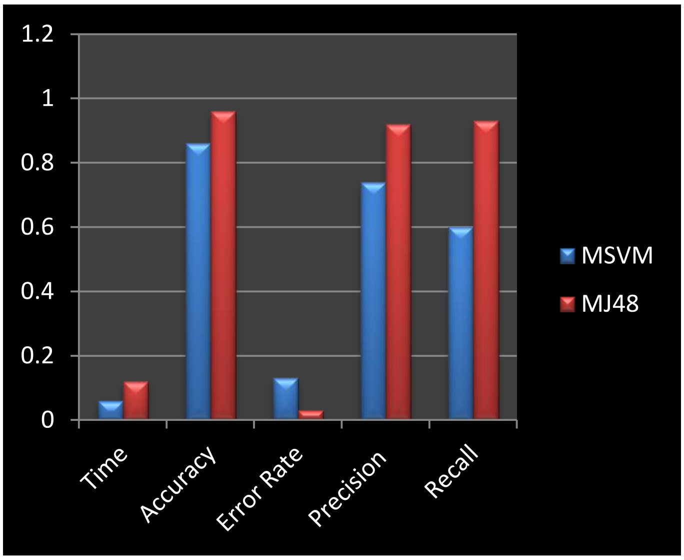

**Figure 6: Comparative Analyses of MSVM and MJ48 Classifier Model**

## VII. CONCLUSION

This research has developed a framework for automatic detection of GDM using hybrid classifier models. The data set has been pre-processed and applied with two proposed models namely MSVM and MJ48. MSVM is the enhanced version of SVM and it provides high accuracy than SVM. MSVM modifies the kernel parameter values as C = 5.0 and $\gamma$ = 1.0. The second proposed classifier model (MJ48) is the modified version of J48 algorithm. It produces result as tree based structure. In MJ48, for boosting the pruning task, the values of the parameters such as Confidence Factor, MinNumobj, Numfolds and Seed are modified as 0.25, 2, 3 and 1 respectively. Based on the parameters such as accuracy, error rate, true positive & false positive ratio, precision, recall, Kappa statistic, Mean absolute error, Root Mean squared error and Relative absolute error, the performance of the proposed models are evaluated and the result shows that the classifier model MJ48 provides very high accuracy of 96.84% with a low error rate of 0.03.

## REFERENCES

1. S. Saradha, Dr. P. Sujatha, "Prediction of gestational diabetes diagnosis using SVM and J48 classifier model", International of Engineering & Technology, Volume 7, Issue 2.21, (2018), pp.323 – 326.
2. G. Thailambal, R. Subramani, S. Saradha, "DRUGS USAGE PREDICTION IN WEKA TOOL USING C4.5 CLASSIFICATION ALGORITHM", International Journal of Pure and Applied Mathematics, Volume 119, No.15, 2018, pp.3633 – 3642.
3. Pavel A, Rudenko, Kirill V, Pozhar, Evgeniia L, Litinskaia, Angelina N. Zhigaylo, " Development of the Short-term Blood Glucose Prediction Algorithm for Using in Closed-loop Insulin Therapy Device", National Research University of Electronic Technology, Moscow, Russia, 978-1-5386-4340-2, 2018.

5. Sarah Ali Siddiqui, Yuan Zhang, Senior Member, Jaime Lloret, IEEE, Houbing Song, and Zoran Obradovic, ”Pain-free Blood Glucose Monitoring Using Wearable Sensors: Recent Advancements and Future Prospects”, IEEE REVIEWS IN BIOMEDICAL ENGINEERING, 1937-3333, 2018.
6. S. Saradha, Dr. P.Sujatha, “ A Performance Evaluation of Classification Algorithms For Diagnosing Gestational Diabetes”, Journal of Advanced Research in Dynamical & Control Systems, 04-Special Issue, June 2017, ISSN: 1943 -023X.
7. Sarangam Kodati, Dr. R P. Singh, ”Comparative Performance Analysis of Different Data Mining Techniques and tools using in Diabetic Disease”, International Journal of Computer Engineering In Research Trends, Volume 4, Issue 12, December - 2017, pp. 556-561.
8. Simon Fong, Jinan Fiaidhi, Sabah Mohammed, Luiz A.M, Moutinho, ”Managing Diabetes Therapy through Datastream Mining”, Published by the IEEE Computer Society 1520-9202, September/October 2017.
9. Ina Maryani and Dwiza Riana, ”Clustering and Profiling of Customers Using RFM For Customer Relationship Management Recommendations”, IEEE **Conference Location:** Denpasar, Indonesia, 17316174, Aug. 2017, 10.1109/CITSM.2017.8089258
10. Miss. N. Vijayalakshmi, Miss. T. Jenifer, ”An Analysis Of Risk Factors For Diabetes Using Data Mining Approach”, International Journal of Computer Science and Mobile Computing, Volume 6, Issue 7, July 2017, pp. 166-172.
11. S. Umadevi, Dr. K. S. Jeen Marseline, “A Survey on Data Mining Classification Algorithms”, International Conference on Signal Processing and Communication ICSPC’17, 28th & 29th July 2017.
12. P. Suresh Kumar and V. Umatejaswi, ”Diagnosing Diabetes using Data Mining Techniques”, International Journal of Scientific and Research Publications, Volume 7, Issue 6, ISSN 2250-3153, June 2017.
13. Ruhi Dubey, Rajni Ranjan Singh Makwana, ”Comparative Analysis of Computer Assisted Valuation of Descriptive Answers using WEKA with different classification algorithms”, SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE), Volume 4, Issue 6, June 2017.
14. S. Kavipriya, Dr. T. Deepa, ”Analyzing the Risk Factors in Gestational Diabetes Mellitus Patients Using Data Mining Rules”, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), Website: www.ijircce.com, Volume 5, Issue 3, March 2017.
15. Suman Morampudi, “The Challenges and Recommendations for Gestational Diabetes Mellitus care in India: A Review”, Front Endocrinol (Lausanne), March 2017.
16. S. Saradha, Dr. P. Sujatha, “Analysis and Significance Study of Clustering Techniques”, IJETER, Volume 4,Issue 9, September(2016), ISSN 2454-6410,pp.31-33.
17. Vimala vinnarasi. A, “Data Mining Techniques: Contemporary Amalgam System to Predict Diabetes.”, IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, ISSN: 2278-8727, Volume 18, Issue 4, Jul.-Aug, 2016, pp.57-60. www.iosrjournals.org.
18. A.A. Ojugo, A.O. Eboka., R.E. Yoro, M.O. Yerokun and F.N. Efozia, ”Hybrid Model for Early Diabetes Diagnosis”, 2015 Second International Conference on Mathematics and Computers in Sciences and in Industry, IEEE, 978-1-4799-8673-6, DOI 10.1109/MCSI.2015.35, 2015.
19. Mani Butwall, Shraddha Kumar, “A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier”, International Journal of Computer Applications (0975 – 8887) Volume 120, No.8, June 2015.
20. D.Sheila Freeda, Dr. Lilly Florence, ”Improving Performance of Diagnosis System for Diabetes Using Data Mining Techniques”, Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com.
21. Dr. P. Sujatha, S. Saradha “A Study of Data mining Concepts and Techniques”, International Journal of Applied Engineering Research (IJAER), ISSN 0973 - 4562, Vol. 9, No.27 (2014), pp.9648-9651.