Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach

Himansu Das, Bighnaraj Naik and H. S. Behera

Abstract The diabetes mellitus disease (DMD) commonly referred as diabetes is a significant public health problem. Predicting the disease at the early stage can save the valuable human resource. Voluminous datasets are available in various medical data repositories in the form of clinical patient records and pathological test reports which can be used for real-world applications to disclose the hidden knowledge. Various data mining (DM) methods can be applied to these datasets, stored in data warehouses for predicting DMD. The aim of this research is to predict diabetes based on some of the DM techniques like classification and clustering. Out of which, classification is one of the most suitable methods for predicting diabetes. In this study, J48 and Naïve Bayesian techniques are used for the early detection of diabetes. This research will help to propose a quicker and more efficient technique for diagnosis of disease, leading to timely and proper treatment of patients. We have also proposed a model and elaborated it step-by-step, in order to make medical practitioner to explore and to understand the discovered rules better. The study also shows the algorithm generated on the dataset collected from college medical hospital as well as from online repository. In the end, an article also outlines how an intelligent diagnostic system works. A clinical trial of this proposed method involves local patients, which is still continuing and requires longer research and experimentation.

Keywords Diabetes mellitus disease (DMD) · Data mining (DM)
J48 · Naïve Bayesian

H. Das (✉) · H. S. Behera
Department of Information Technology, Veer Surendra Sai University of Technology,
Burla, Sambalpur 768018, Odisha, India
e-mail: das.himansu2007@gmail.com

H. S. Behera
e-mail: mailtohsbehera@gmail.com

B. Naik
Department of Computer Application, Veer Surendra Sai University of Technology,
Burla, Sambalpur 768018, Odisha, India
e-mail: mailtobnaik@gmail.com

P. K. Pattnaik et al. (eds.), *Progress in Computing, Analytics and Networking*,
Advances in Intelligent Systems and Computing 710,
https://doi.org/10.1007/978-981-10-7871-2_52

1 Introduction

Out of many chronic diseases, diabetes is one of them. It occurs mainly when the pancreas is unable to produce the desired amount of insulin required for a human body or when the human body cannot effectively manage the produced insulin. When taken as a whole, the risk of dying among the people with diabetes is at least double the hazard of their peers without diabetes. As per the prediction of World Health Organization (WHO), diabetes will be one of the major leading causes of death in 2030 and death rate will double between 2005 and 2030. It could also be described as the situation in which the body is unable to process properly the food for utilization as energy. The majority of the food consumed by us is turned into glucose, which is further used as energy by us when required. The pancreas, an organ of our body which lies near the stomach of a human body, produces a special type of hormone called insulin helps glucose to get into the cells of our body. The persons affected by diabetes, the body of the affected persons are either do not produce enough insulin or unable to consume its own insulin as well. This leads to increase the sugar level in our blood. For this cause of increasing the sugar in the blood, people call diabetes as 'sugar'. Several symptoms are normally found in the persons affected by this disease. The main symptoms are frequent urination, feeling pain in the muscles, increased hunger, and thirst. It needs early detection of the disease. It can cause many severe complications if not treated at the early stage. Short-term complications include diabetic ketoacidosis, nonketotic hyperosmolar coma, or death. Major chronic complications may arise like dysfunction of heart which leads to stroke, foot ulcers, chronic kidney failure, damage to the retinas of the eyes, nerves, and teeth. The diabetes is mainly categorized in three types [1]. First, the Type 1 DMD occurs due to the failure of pancreas to produce enough insulin. The cause still remains unknown. Second, Type 2 DMD begins with insulin resistance, a situation in which cells will be unsuccessful to respond to insulin appropriately. With the progress of diabetes, be short of insulin may also develop. The common causes may be due to excessive body weight and insufficient exercise required in the changed lifestyle. And Third, gestational diabetes, which occurs at the time of pregnancy of a women without a prior history of diabetes suddenly increases high blood-sugar levels. Advancement in the field of computer science and high-performance computing has benefited almost all disciplines including medical science in finding better results over traditional practical solutions. Many tools have been developed for effective analysis of image processing and effective analyzing by applying data mining techniques in order to help clinicians in making better decisions for the diagnosis of the diseases of patients. Nowadays, data mining plays a vital role and becomes an essential methodology for medical diagnostics. Data mining helps in finding the hidden pattern lies in the pathological data, large-scale medical images and the daily records to understand more clearly about the hidden relationships between diagnostic features of different patient groups [2–4]. Nowadays, data mining has been used extensively in the areas of science and engineering, such as genetics, bioinformatics, medicine, and

education [5–9]. The primary objective of our research is to find whether a patient is affected by diabetic or not. In order to accomplish this objective, very popular data mining algorithms like J48 and Naïve Bayesian are used. By using this algorithm, it becomes easy to understand the whole process [10–13]. These algorithms help to analyze the disease from all the aspects.

The organization of the remaining section of this paper is as follows. Section 2 discusses the basic concepts to understand our work. The related works are represented in Sect. 3. Section 4 introduces our proposed approach for detecting diabetes mellitus disease. Section 5 analyzes the experimental results of the collected dataset, and Sect. 6 describes the comparison of the techniques after applying proposed approach. Section 7 concludes the paper.

2 Basic Concepts

This section presents some of the basic algorithmic concepts which are required to understand the proposed approach.

2.1 Naive Bayesian Algorithm

It is a supervised learning optimal classifier algorithm that is based on the concept of Bayes' theorem. This model is easy to construct, with easy iterative parameter assessment which makes it mostly valuable for very huge datasets. This classifier is used to convert the prior probability into posterior probability by using likely used values. This Naive Bayes' theorem is also often used to carry out complex classification tasks.

This theorem provides a technique to calculate the posterior probability P(X|Y), from P(X), P(Y), and P(Y|X). This Naive Bayes classifier assumes that the outcome of the cost of a predictor (Y) on a given class (X) is independent of the costs of other predictors. This hypothesis is called class conditional independence. The Bayes' theorem is represented in Eq. 1.

$$P(\mathrm{X}|\mathrm{Y}) = (\mathrm{P}(\mathrm{Y}|\mathrm{X})\mathrm{P}(\mathrm{X}))/\mathrm{P}(\mathrm{Y}) \tag{1}$$

Here, P(X|Y) is the posterior probability of X conditioned on Y, P(X) is the prior probability or apriori probability of X, similarly P(Y|X) is the probability of Y conditioned on X and P(Y) is defined as the prior probability.

2.2 J48 Algorithm

The J48 decision tree algorithm is used for classification. It uses each and every phase of the data attribute which is divided into smaller subclasses to establish on a decision. The J48 algorithm inspects the normalized information gain that outcomes dividing the data by selecting an attribute. In generally, the higher the normalized information gain of attributes is used to build the decision then this algorithm recurs the lesser reduct of elements. The J48 algorithm selects the particular attributes, dissimilar element values, and lost feature values of the information. If all the instances of the subset are considered with the same set, then after splitting the instances fit into the identical set. All attributes in the sequence will be measured, and the expansion in information will be chosen for the confirm on the attribute. The suitable element will be recognized to resultant from the recent classification constraint.

3 Related Work

Not much work has been done in the field of diabetes mellitus disease (DMD). Here, some of the existing work is represented as follows. Sankaranarayanan and Pramananda Perumal [14] focused their study on two classification methods, i.e., rule classification and decision trees. The decision tree was based on the algorithm from the rule classifier. The attributes specified after the physical examination research were code, sex, and age. The concept of making an algorithm by considering all the attributes was taken from their study and rule classification method. They even gave the concept of data warehouse which could be used for further examination of patients and their classification which is a considerable point for our research. Further research of Iyer et al. [15] compared two of the major algorithms of data mining techniques J48 decision trees and Naive Bayesian algorithm. From the results obtained, both the methods have a comparatively small difference in error rate, though the percentage split of 70:30 for Naïve Bayes technique gives the least error rate as compared to other J48 implementations. According to their algorithms, Naive Bayes had a lesser error rate than J48 decision tree, and hence, it was considered better. These algorithms were more accurate and specific for classification. These algorithms gave faster results than those used in our paper hence could be used for future modification.

In some other research of Velu and Kashwan [16], they worked on Pima Indian Diabetes (PID) dataset. They used three techniques such as genetic algorithm (GA), EM algorithm, and h-means with clustering. The result stated that the double crossover genetics process and h-means + based techniques were superior on performance comparison. They performed simulation on WEKA software for three

models to test the classification. For this simulation tests, they have taken seven attributes with 768 instances of diabetic patients gathered from the different hospitals. Their results showed that out of 768 instances, 500 patients were experienced with negative and 268 patients were experienced as positive. Correlation coefficient was determined to be 0.96 which specifies that there was a strong relationship between these two samples. They basically classified diabetic patients as positive and negative on the basis of seven attributes by using the above algorithms. This classification technique is used as a reference in our work.

Motka et al. [17] proposed four different approaches for classifying the disease into two main classes: diabetic and non-diabetic. They used the techniques PCA with ANFIS, ANFIS, neural detworks, and PCA with neural networks. They observed that after combining PCA with neural network for classification gives the better accuracy. They used soft computing techniques for classification and MATLAB GUI. Their classification was proper and simple. Rajesh and Sangeetha [18] applied many classification algorithms on diabetes dataset and analyzed the performance of those algorithms. A classification rate of 91% was obtained for the C4.5 algorithm. Future enhancement of this work includes modification of the C4.5 algorithms to advance the classification rate to achieve greater accuracy in classification. C4.5 is considered the best of all the algorithms for classification. Patil and Joshi [19] in their study gave a certain set of rules using association rule data mining algorithms, i.e., apriori algorithm which is further used in this study also. They implemented it in WEKA and generated those top ten powerful rules for diabetes = 'yes'. Whereas diabetes is equal = 'no'. These rules had their own support and confidence. In this paper, Apriori algorithm for classification of DMD is implemented and considered their rules also.

4 Proposed Model

In this section, proposed method for determining diabetes is presented. This system saves the time of the doctors as well as of patients by generating the report through its data repositories. This generator determines whether the patient is diabetic or not. Figure 1 represents the block diagram of the proposed framework.

The proposed work can be vitalized as presented below:

Patients: They are the main focus point of our system. They are responsible for providing all the data for the repository. With their help, the process of collecting data becomes very easy.

Data collection from patients: In order to collect data from patients, they are provided with the set of questions. Those questions include name, age, sex, blood sugar level, and plasma glucose concentration which is a 2 hours in an oral glucose tolerance test, etc.

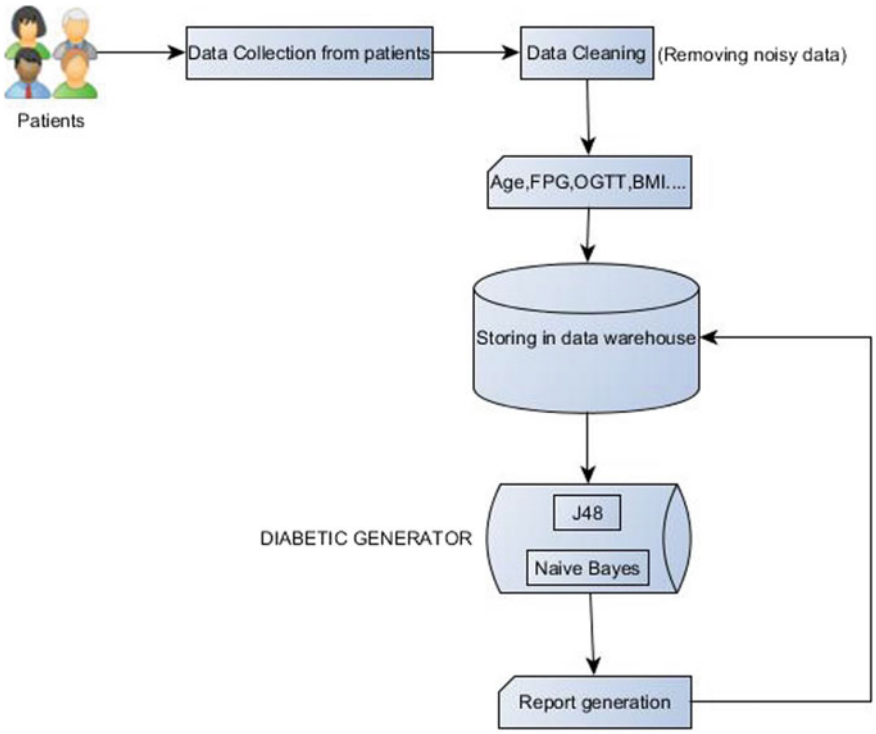

Fig. 1 Block diagram of the proposed model

Data cleaning step is very decisive. Here, unnecessary data are removed and only important data are sent for further implementation, for example, sex, age, oral glucose tolerance test (OGTT). Next step is storing in the data warehouse, where the data are stored in the data warehouse. This is done so to maintain all the records about patients which could be referred whenever required. Finally, in diabetes generator step which generates a report after processing all the data by specifying whether the patient is diabetic or non-diabetic.

5 Results and Discussion

We have collected 200 data by preparing a questionnaire given in Fig. 2, from a local medical college hospital out of which 60% is used to training purposes of the model and the rest 40% of the collected data is used to test in our model. The data are collected first and then preprocessed to fill the missing values. Some redundant data are also omitted. For the future use, the data are stored in the warehouse.

1. Name :

2. Age:

3. Gender : Male ☐ Female ☐ (*tick one*)

4. Occupancy : (WORKING/NOTWORKING)

Instructions:

This questionnaire will help us to find out the summary of occurrence of diabetes in people of various age groups. Each item in this questionnaire is describing a specific attribute of the diabetes disease. We want your opinion as how likely is the occurrence of diabetes with the given set of attributes. Given the actual choices that you face, tick the choice that you will prefer. There are six questions in all.

1. Age (Years)

i. 21-35 ☐ ii. 36-50 ☐ iii. 51-65 ☐ iv. 66-71 ☐

2. Body Mass Index (kg/m^2) = *Weight:* , *Height*:

i. 22-24.9 ☐ ii. 25-30 ☐ iii. 31-40 ☐ iv. 41-68 ☐

3. Diastolic Blood Pressure (mm/Hg) (*Lower range pressure*)

i. 50-70 ☐ ii. 71-85 ☐ iii. 86-100 ☐ iv. 101 -114 ☐

4. Diabetes Pedigree Function *(Hereditary)*

i. Yes ☐ ii. No ☐

5. Plasma Glucose Concentration (mg/dl)

i. 78-108 (low) ☐ ii. 109- 134 (intermediate) ☐

iii. 135-170 (high) ☐ iv. 170- 199 (Very High) ☐

6. Are you suffering from Diabetes?

i. Yes ☐ ii. No ☐

7. Type of Diabetes :

Fig. 2 Questionnaire to collect the data from the patients

By using the diabetes generator, the data are classified. Results obtained from this model are as follows:

In this dataset, total eight numbers of attributes are used, namely age (years), plasma glucose concentration test, triceps skin fold thickness (mm), diastolic blood pressure (mm Hg), 2 hours serum insulin (mu u/ml), diabetes pedigree function, body mass index (weight in kg/(height in m)^2), and class variable.

Each attribute used has its own role in predicting whether the patient is diabetic or non-diabetic. Using WEKA software, the graph is generated in Fig. 3 which describes the possibility of having diabetes and not having. Red color shows the test is positive, i.e., a person is having diabetes and blue shows tested negative.

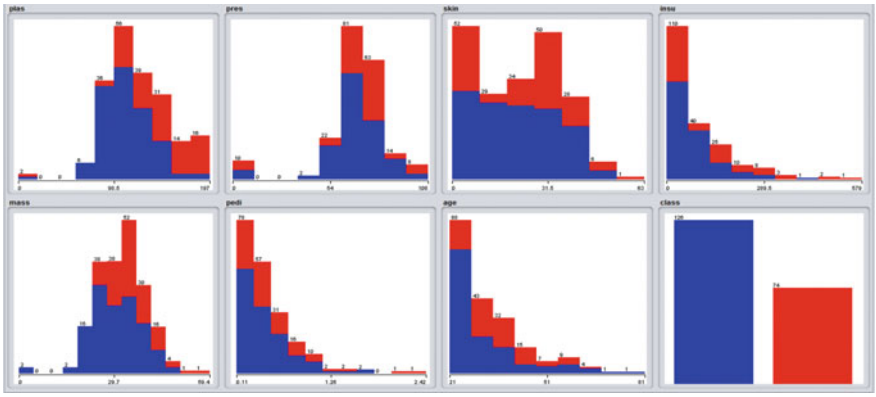

Fig. 3 Graphical representation of all the attributes of dataset

Table 1 Description of output after applying Naive Bayers theorem

Tested negative	Tested positive	
107	19	Tested negative
36	38	Tested positive

At first, Naive Bayes theorem is applied in WEKA software. In the output, we observed that 145 instances are correct and 55 instances are incorrect. Through this, we got the exact statistical measures; few of them are as follows (Table 1):

i. The Mean absolute error—0.3142
ii. Relative absolute error—67.3187%
iii. Root relative squared error—92.2618%
iv. Root mean squared error—0.4456
v. Kappa statistics—0.3808.

Next, the J48 method is applied which shows 139 instances to be correct and 61 instances to be incorrect out of total 200 instances. The statistical measures are as follows (Table 2 and Fig. 4):

i. Mean absolute error—0.3822
ii. Relative absolute error—81.9009%
iii. Root relative squared error—100.5779%
iv. Root mean squared error—0.4857
v. Kappa statistics—0.3051.

Table 2 Description of output after applying J48 theorem

Tested negative	Tested positive	
106	20	Tested negative
41	33	Tested positive

```
J48 pruned tree
------------------

plas <= 139
|   mass <= 26.2: tested_negative (44.0/1.0)
|   mass > 26.2
|   |   plas <= 94: tested_negative (23.0/1.0)
|   |   plas > 94
|   |   |   age <= 47
|   |   |   |   age <= 34: tested_negative (54.0/17.0)
|   |   |   |   age > 34
|   |   |   |   |   pres <= 70
|   |   |   |   |   |   plas <= 118: tested_negative (3.0)
|   |   |   |   |   |   plas > 118: tested_positive (3.0)
|   |   |   |   |   pres > 70: tested_positive (10.0)
|   |   |   age > 47: tested_negative (6.0)
plas > 139: tested_positive (57.0/15.0)

Number of Leaves  :     8

Size of the tree :      15
```

Fig. 4 J48 pruned tree that was generated by WEKA

6 Comparison

In this section, both the algorithms are used to test the diabetes dataset and the results are described. Table 3 compares both the methods and could easily conclude that Naive Bayes theorem is better than the J48 as the time to build the model is less.

Time to build the module and appropriately classified instances are higher when Naive Bayes algorithm is used, and classification accuracy is also higher in Naive Bayes algorithm than that of J48 algorithm. The above results also that on diabetes dataset Naïve Bayes performs much more better than of J48. Figure 5 makes it clearer to compare the performance of both the methods.

Table 3 Comparison between J48 and Naive Bayesian theorem

Evaluation criteria	J48	Naïve Bayesian
Time to build module (in seconds)	0.08	0
Correct classified instances	139	145
Incorrect classified instances	61	55
Prediction accuracy	69.5	72.5

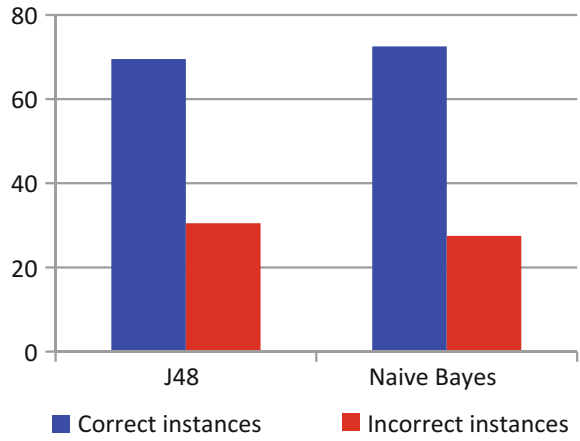

Fig. 5 Performance of J48 with Naive Bayes

7 Conclusion

Data mining is an integrated field with the vast variety of techniques from several fields. It is a combination of machine learning, statistics, pattern recognition, and artificial intelligence systems for analysis of huge amount of data to discover the hidden patterns in the data. Nowadays data mining techniques are applied in medical sciences for decision making. It also plays an essential role in diabetes dataset to expose and uncover the hidden knowledge from a massive amount of unused diabetes data that will significantly assist to progress the quality treatment for the patients suffering from diabetes. In this research, the classification techniques used for predicting diabetes in patients are J48 and Naive Bayes theorem. Naive Bayesian theorem performs more efficiently and effectively as compared to J48. Whether it is time to build the model or identifying correct instances or accuracy, Naive Bayes always proved its productivity.

In future work are planned to propose more complex and integrated model by hybridizing machine learning techniques which will be able to predict all types of diabetes. Further, it will also include collecting data from different local areas of the country in order to get more précised and accurate data which will result in more précised and accurate outcome.

References

1. American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 37, no. Supplement 1 (2014): S81–S90.
2. Thirumal, P. C., and N. Nagarajan. "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study." ARPN Journal of Engineering and Applied Science 10, no. 1 (2015).

3. Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients." International Journal of Engineering and Advanced Technology 1, no. 3 (2012): 147–151.
4. Kaur, Gaganjot, and Amit Chhabra. "Improved J48 classification algorithm for the prediction of diabetes." International Journal of Computer Applications 98, no. 22 (2014).
5. Daghistani, Tahani, and Riyad Alshammari. "Diagnosis of Diabetes by Applying Data Mining Classification Techniques." International Journal of Advanced Computer Science and Applications (IJACSA) 7, no. 7 (2016): 329–332.
6. Marinov, Miroslav, Abu Saleh Mohammad Mosa, Illhoi Yoo, and Suzanne Austin Boren. "Data-mining technologies for diabetes: a systematic review." Journal of diabetes science and technology 5, no. 6 (2011): 1549–1556.
7. Shivakumar, B. L., and S. Alby. "A survey on data-mining technologies for prediction and diagnosis of diabetes." In Intelligent Computing Applications (ICICA), 2014 International Conference, pp. 167–173. IEEE, 2014.
8. Christobel, Y. Angeline, and P. Sivaprakasam. "A New Classwise k Nearest Neighbor (CKNN) method for the classification of diabetes dataset." International Journal of Engineering and Advanced Technology 2, no. 3 (2013): 396–200.
9. Das, Himansu, Ajay Kumar Jena, Janmenjoy Nayak, Bighnaraj Naik, and H. S. Behera. "A novel PSO based back propagation learning-MLP (PSO-BP-MLP) for classification." In Computational Intelligence in Data Mining-Volume 2, pp. 461–471. Springer, New Delhi, (2015).
10. Amit kumar Dewangan, Pragati Agrawal.: Classification of Diabetes Mellitus Using Machine Learning Techniques. Vol. 2, 5 (2015).
11. Srikanth, Panigrahi, and Dharmaiah Deverapalli. "A critical study of classification algorithms using diabetes diagnosis." In Advanced Computing (IACC), 2016 IEEE 6th International Conference on, pp. 245–249. IEEE, 2016.
12. Saravananathan, K., and T. Velmurugan. "Analyzing Diabetic Data using Classification Algorithms in Data Mining." Indian Journal of Science and Technology 9, no. 43 (2016).
13. Saxena, Krati, Zubair Khan, and Shefali Singh. "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm." International Journal of Computer Science Trends and Technology (IJCST) (2014).
14. Sankaranarayanan, Sriram, and T. Pramananda Perumal. "A predictive approach for diabetes mellitus disease through data mining technologies." In Computing and Communication Technologies (WCCCT), 2014 World Congress on, pp. 231–233. IEEE, 2014.
15. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." arXiv preprint arXiv:1502.03774 (2015).
16. Velu, C. M., and K. R. Kashwan. "Visual data mining techniques for classification of diabetic patients." In Advance Computing Conference (IACC), 2013 IEEE 3rd International, pp. 1070–1075. IEEE, 2013.
17. Motka, Rakesh, Viral Parmarl, Balbindra Kumar, and A. R. Verma. "Diabetes mellitus forecast using different data mining techniques." In Computer and Communication Technology (ICCCT), 4th International Conference on, pp. 99–103. IEEE, 2013.
18. Rajesh, K., and V. Sangeetha. "Application of data mining methods and techniques for diabetes diagnosis." International Journal of Engineering and Innovative Technology (IJEIT) 2, no. 3 (2012).
19. B. M. Patil, R. C. Joshi, Durga Toshniwal.: Association rule for classification of type-2 diabetic patients (2010).
20. Vijayan, Veena, and Aswathy Ravikumar. "Study of data mining algorithms for prediction and diagnosis of diabetes mellitus." International journal of computer applications 95, no. 17 (2014).