Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2021.020401
ISSN 2790-0983 Vol. 2, Issue 4: 1-9

Diabetes Prediction Based on Random Forest Algorithm

Jie Cai*

Guangzhou City Construction College, Guangzhou, China

**corresponding author*

Keywords: Random Forest Algorithm, Diabetes Prediction, Feature Selection, Data Analysis

Abstract: With the development of artificial intelligence, society will definitely become smarter in the future. More and more intelligent and convenient products will appear in everyone's life. For the medical industry, the huge amount of data generated is a valuable asset. How to discover the patterns of diseases in these data and improve the efficiency and accuracy of disease prediction through scientific means has become a hot issue for research nowadays. The main objective of this paper is to investigate the prediction of diabetes based on the RFA. In this paper, the RF algorithm (RFA) is used to analyse diabetes data, and the method is proposed based on the study of different DP models to improve the effectiveness of diabetes prediction (DP). Based on the RF-based ranking of feature importance, the features are added to the set of features to be evaluated in order of importance; for the proposed model is unable to perform effective feature selection, greedy feature selection based on RF is incorporated to further improve the accuracy of the DP model. The system was tested and experimented with, and the results showed that the prototype system of the DP model can effectively achieve the prediction function of diabetes.

1. Introduction

As living conditions in the country improve and the problem of an ageing population increases, this has led to an exponential increase in the number of people with diabetes. However, current measures for the prevention and diagnosis of diabetes are not well suited to the needs of the diabetic population. At the same time, the development of information technology has led to the gradual accumulation of a large amount of diabetes-related data in the medical industry. How to make use of these diabetes data, discover the hidden patterns in them and further use scientific means to predict diabetes is currently a hot research issue in the field of diabetes [1-2].

The research shows that there are only a few papers in China that use RFA to process some modern medical data to analyse the prevalence factors of diabetes and design relevant early warning

systems. In China, early warning models for diabetes complications have been constructed and compared using neural network algorithms based on MATLAB and SPSS [3]; the accuracy and superiority of early warning models constructed by improved neighbourhood rough set algorithms have been tested in Python [4]; and three different algorithms have been used to build early warning models to analyse the prevalence of diabetes. The model based on BP artificial neural network was found to be more efficient than SUV and integrated learning algorithms [5]. Overseas, research on early warning algorithms for diabetes has been conducted for a longer period of time than in China. For example, Nora et al. used a regression model to predict the risk of diabetes complications within a certain time frame [6], using data from some conventional physical indicators and a medical indicator, the glycosylated haemoglobin ratio, and then concluded that the model performed well by looking at the curve distribution. From the above analysis, it is clear that there is a certain practical basis for the application of machine learning to DP at home and abroad, and the future development prospects are even more promising.

In this paper, we investigate the prediction of diabetes based on the RFA. Firstly, the RFA is used to analyse diabetes data, and the method is proposed based on the study of different DP models, which improves the effectiveness of DP. Secondly, based on the RF ranking of feature importance, the features are added to the set of features to be evaluated in order of importance, and the best combination of features is found by applying the idea of greed to keep the features that make the evaluation model more effective. The final selection of diabetes features is achieved. The proposed model is unable to perform effective feature selection because of the pursuit of high accuracy of complex models in the current study of DP models. The proposed model incorporates greedy feature selection based on RF to further improve the accuracy of DP models. Finally, the application scenarios of the proposed DP model are analysed and a prototype system of the DP model is designed for the above proposed DP model.

2. Design Research

2.1. Problems and Development Trends

Through the above analysis of the current status of research on diabetes early warning systems at home and abroad, it can be seen that research on early warning systems for diabetes through medical data related to Western medical diagnosis is not common. Therefore, in the diabetes early warning system designed and implemented in this paper, the RF classification algorithm was used to analyse the relevant data and find that there is a potential relationship between the values of various attributes in the diabetes prevalence factors and having diabetes, and use this knowledge as a basis for determining whether or not one is likely to have diabetes [7-8].

The complexity of the medical data information covered and the high data requirements of the algorithm require the extraction of useful data information and a high level of accuracy of the decision data. It is therefore necessary to collect advice from experienced clinicians, apply relevant machine learning as well as statistical knowledge to synthesise innovations and train algorithms through data analysis and models to obtain results with a high accuracy rate. With the development of the Python language, the widespread use of the RFA, and the continuous improvement of the Django framework, the development of combining computer science with the medical field will become more extensive and deeper, thus bringing greater benefits to human society [9-10].

2.2. Overview of System Requirements

As most people have poor self-control, uncontrolled diet, and also lack of awareness of diabetes prevention, and in addition, through the introduction of this paper, we learn that the number of

low-age diabetic patients shows a continuous increase, therefore, the design and implementation of this early warning system for diabetes has a practical significance, through the promotion of this system can let more users understand their probability of diabetes, so as to achieve the role of early warning [11-12].

1) Functional requirements analysis

The overall functional diagram of the RF-based diabetes early warning system is shown in Figure 1.

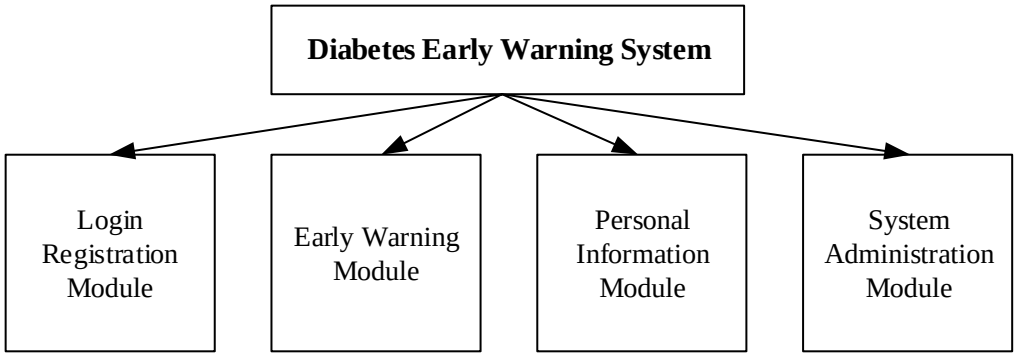

Figure 1. Overall functional diagram of the early warning system

The system is presented in the form of a web page to achieve the system interface. The most important function is the early warning module for diabetes warning. In addition, in order to optimise the system and enrich the user experience, three auxiliary modules have been added, including login and registration, personal information and system management.

2) Non-functional requirements

The analysis of non-functional requirements cannot be overlooked in the system design process, and echoes the functional requirements that must be considered in the system development process. It usually includes the technical and business performance, reliability, maintainability and scalability of the system.

(1) Performance: The core work of the system is the creation of the early warning model, so the performance of the early warning model mainly consists of the analysis of the length of the model training time and the length of the waiting time for the early warning module to give the results after the prediction. Generally speaking, model training time is generally determined by the size of the data volume and the strength of the algorithm, and the length of time varies, while the RFA we use has a fast training speed, coupled with the relatively small amount of data we currently collect, so results can be obtained in a short time. As for the response time for waiting for results, generally speaking, the response time is in the range of 2-5 seconds is the best, so it is important to reduce the response time of the system to within 5 seconds.

(2) Maintainability: The database storage design should be reasonable and clear, and the system should be easy to maintain.

(3) Reliability: It should be ensured that the system does not crash under operational conditions.

(4) Expandability: In the process of using the system, as the number of users increases, user information will also gradually increase, so it is necessary to retrain the early warning model, so as to improve the accuracy of the analysis results of the early warning system.

3) Feasibility analysis

The Django framework and RFs have been used in other areas and have a proven track record, so it is possible to use them in this paper. The system is developed on a Windows operating system, and the operating conditions required for the system in this paper are not difficult to achieve, as long as some minimal hardware and software configurations are available to keep the system

running [13-14]. It is possible to rely on existing computer equipment in the laboratory, and the installation packages that need to be used are easy to obtain, with no additional financial requirements.

2.3. Information Gain Algorithm

The training dataset D, |D| is the sample size, i.e. the number of elements contained in the sample D. The class K is denoted by Ck. |Ck| is the number of samples in the subsample, and the sum of |Ck| to |D|, k=1,2 ..... , according to some element A, D is divided into n subsets D1, D2 ..... Dn, |Di| is the number of samples of Di, the sum of |Di| is |D|, i=1,2, ... , in addition, the set of samples belonging to Ck of Di is Dik, the intersection set, where |Dik| is the number of samples of Dik, and the algorithm is shown in Table 1.

Table 1. Information gain algorithm process table

Inputs.	D,A
Outputs.	Information gain g(D,A)

D The empirical entropy H(D) probability is calculated based on classical probability, because the total number of training sets is |D|, so the number of any one classification can be set to |Ck|, and the probability of this classification is expressed as: |Ck|/D|. The formula is as follows.

$$H(D) = -\sum_{k=1}^{k} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \tag{1}$$

The empirical conditional entropy H(D|A) of choice A. This probability is also calculated based on classical probabilities. Since |Di| is the number of samples for some classification of the selected feature element, then |Di|/|D| is the probability of classifying the selected feature, the following summation can be considered as the entropy of the conditional probability under some category of the selected feature element, i.e. the training set is Di, and Dik can be considered as the number of samples to be classified under some Di condition, i.e. k is some classification, i.e. the entropy of reducing the training set to Di , the formula is as follows.

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{k} \frac{|D_i|}{|D|} \log_2 \frac{|D_{ik}|}{|D|} \tag{2}$$

The information gain g(D,A) is given by

$$g(D,A) = H(D) - H(D|A) \tag{3}$$

The information gain ratio information gain has the problem of selecting features with more values, so a method is proposed to correct this problem, i.e. the information gain ratio. It is calculated as follows.

$$g_R(D,A) = \frac{g(D,A)}{H_A(D)} \tag{4}$$

$$H_A(D) = -\sum_{i=1}^{m} \frac{|D_i|}{|D|} \log \frac{|D_i|}{|D|} \tag{5}$$

m represents the number of values taken for feature A. It is equivalent to multiplying the information gain by the penalty parameter.

3. Experimental Study

3.1. Simple Feature Selection Based on RF

The main idea of simple RF feature selection (SRF) is: first, the data is trained using RF; then, the OOB data classification correctness method of RF is used to quantify and rank the contribution of each feature in the data to the model; then, the top features ranked in order are selected from the features to participate in the training of the external evaluation model, where i=1,2 ,...,m, with m denoting the dimensionality of the data; finally, the set of features with the highest ratings is taken as the result of this feature selection.

In order to reduce the complexity of feature selection, the step size step is increased to control the increase of i. To improve the stability of feature selection, K-fold cross-validation is used in the evaluation.

The flow of the SRF feature selection algorithm is as follows.

Step1: Using the RF dataset, obtain the degree of contribution of each of the dataset to the RF model building, and quantify the ranking.

Step2: Select the top k∗ step features to form a feature subset, where k=1,2,...,m/step. step controls the increase in the number of features selected and is used to reduce the complexity of feature selection. When step=1, k=1, 2,..., m.

Step3: Take the selected subset of features and evaluate them using an external evaluation model.

Step4: The subset of features with the highest ratings is selected for output.

In this paper, a logistic regression model is used as the external evaluation model in the subsequent experiments for the sake of comparison. And a 5-fold cross-validation method is used in building the evaluation model.

3.2. System Testing

The final step in the development and design of the system is system testing, which is a critical step. Through testing, we can understand the possible errors and problems in the operation of the system analyse the specific problems and solve them, so as to achieve the perfection and improvement of the system. Black box experiments are generally used to test the functionality of the system, usually on the external parts of the system, such as testing the login function of the system, which is a test of whether the login interface of the system is normal and whether the login operation can be achieved, without dealing with the internal procedures.

The purpose of testing the system is to check whether the various modules of the system are working properly, whether users can log in and register normally, whether information can be modified, whether diabetes alert analysis can be carried out and the results obtained, and whether the relevant operations of the administrator are carried out normally. If problems are found, the system will be debugged and corrected in time to ensure that all functions of the system are running smoothly.

For testing the stability and operability of the various functions of the RFA-based diabetes warning system in this paper, the test functions of the warning system are shown in Table 2.

Based on the above functional use case analysis, the functionality of each module of the system is then tested and the corresponding test results are obtained.

Table 2. Functional test cases

No.	Test instructions	Functional completeness / data correctness
1	User Login Test	Users who have completed registration can log in normally
		The system administrator can log in normally
2	User registration test	The user can enter the registration page by clicking the registration button
		Fill in email and password, registration is successful
3	Information Management Test	Click on the Personal Centre button to enter the Personal Centre page
		Click on Change Password to change password
4	Early warning module test	Click on the warning analysis, can enter the information filling page
		You can fill in the corresponding information
		Click on Submit
		The results of the alert analysis are displayed
5	System management module test	Click on user management, you can delete users
		The user password can be changed
		Ability to manage user information on analysis results

4. Experiment Analysis

4.1. Comparative Analysis of RF Predictions

To compare the goodness of the models, evaluation metrics were incorporated as a measure. RFW-SVM, MKLSVM, SMOTE+MKLSVM and PSO-RFFW-MKLSVM were re-used for experiments on dataset B. The detailed comparison results are shown in Table 3.

Table 3. Comparison results for RFW-SVM, MKLSVM, SATT+MKLSVM and PSO-RFFW-MKLSVM

	RBF-SVM	MKLSVM	Smote+MKLSWM	PSO-RFFW-MKLSVM
Number of features	7	13	13	7
Classification accuracy %	90.4	88.6	89.3	94.7
Sensitivity%	72.4	75.9	82.8	87.9
Specificity %	95.1	91.9	91.0	96.4
Jorden Index	0.67	0.68	0.74	0.84

As can be seen from Figure 2, the overall classification results of RFW-SVM, MKLSVM and SMOTE+MKLSVM are relatively similar; RFW-SVM improves the classification accuracy due to the removal of redundant features; MKLSVM improves the overall performance of the classifier due to the use of a multi-core kernel function with better mapping performance to analyze the data, but both models for a small number of classes However, neither of these models has a high recognition rate for a few classes; SMOTE+MKLSVM greatly improves the reliability of the model in terms of the overall classification effect; PSO-RFFW-MKLSVM further takes into account the data imbalance caused by the different importance of features, and combines the improved RFA on the basis of SMOTE+MKLSVM, by using the calculated feature importance scores to multi-kernel functions Feature weighting was performed to enhance the influence of strongly correlated features on the classification results, substantially improving the recognition rate of diabetic patients and the reliability of the prediction model. The PSO-RFFW-MKLSVM classification, which combines the advantages of RF and multicore support vector, was found to be the best and could effectively predict the risk of diabetes.

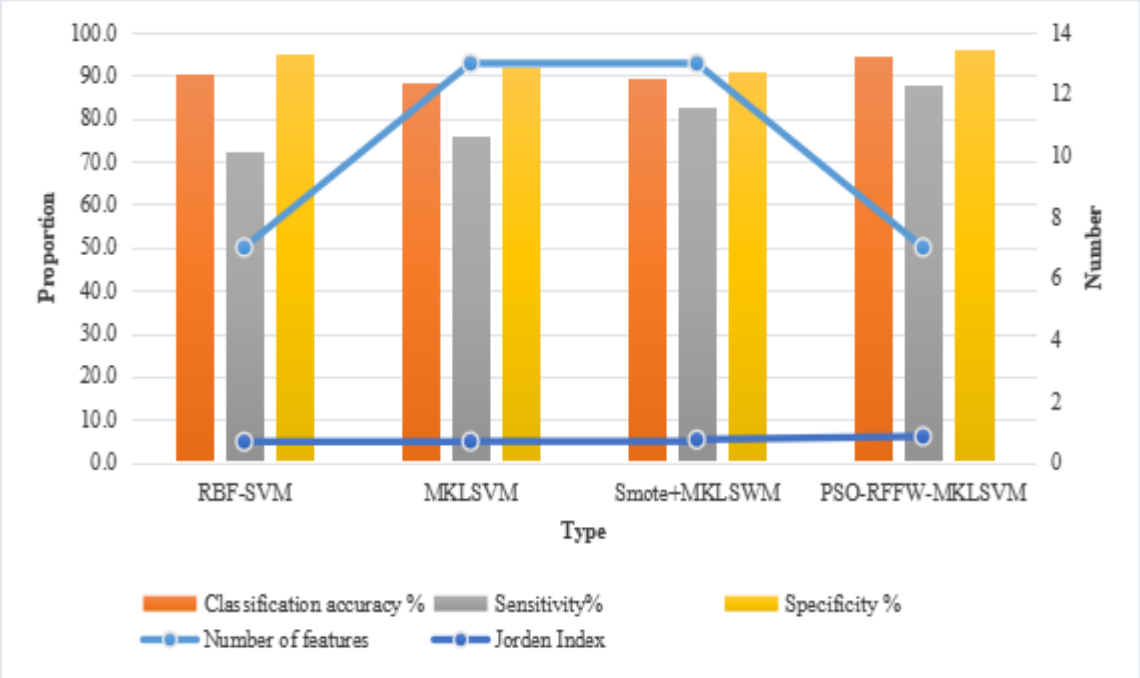

Figure 2. Analysis of the comparative results of RFW-SVM, MKLSVM, SATT+MKLSVM and PSO-RFFW-MKLSVM

4.2. System Testing and Experiments

System testing is the final and most important step involved in system development. Through testing, possible problems in the system development can be well identified and then analysed and corrected to improve the whole system. The purpose of this test is to verify that the various modules of the prototype system are working properly.

Table 4 shows the testing of each functional module.

Table 4. Table of test results for each functional module

Test subjects	Function Description	Interface	Functionality
Login	Ability to log in normally	Friendly	Good
Register	Enables normal registration	Friendly	Good
Permissions Module	Isolated access	None	Good
Information entry	Can enter normally	Friendly	Good
Data processing	Abnormal input can be handled	Friendly	Good
Result prediction	Can give results based on input	Friendly	Good
Record queries	Ability to query historical data	Friendly	Good
Historical statistics	Ability to query historical statistics	Friendly	Good

As shown in Table 4, all modules passed the test requirements well and all modules of this prototype system worked properly.

The following will further analyse the effectiveness of the prototype system by conducting experiments on the prediction module and the information query module. For the prediction module, the experimental verification of a single input message is carried out first, and the results of the prediction are shown in Figure 3.

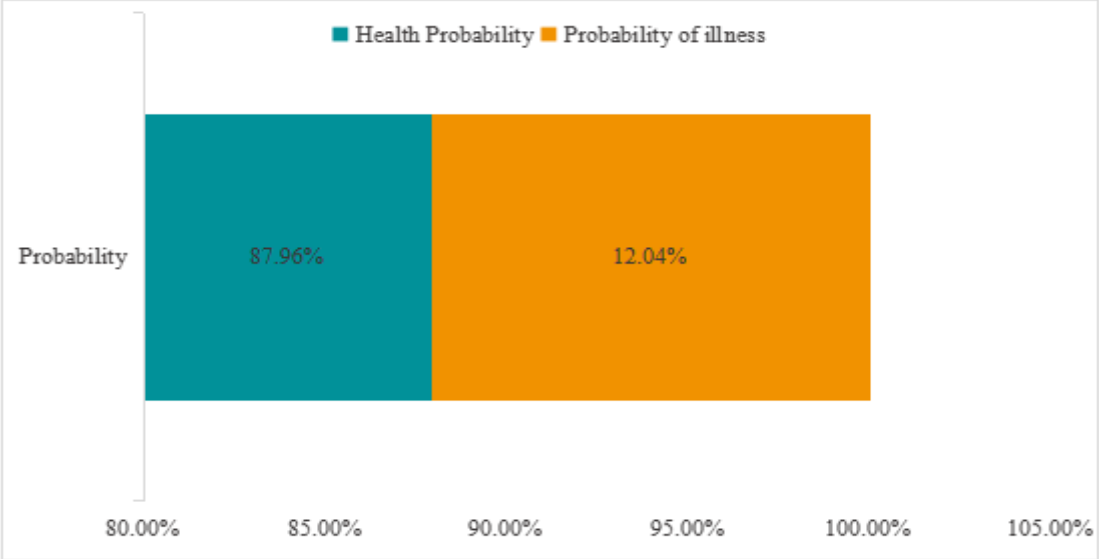

Figure 3. Graph of test results

As can be seen in Figure 3, for the above test data, the prototype DP model system, predicts a 12.04% probability of having the disease and therefore the system predicts it as healthy.

5. Conclusion

In recent years, machine learning has become popular all over the world. There is a very strong potential for the application and development of machine learning in the medical field. As a result, intelligence in diagnosis and treatment is widely proposed. However, due to the great diversity of patient information and the various disease prevalence factors, much of the data information collected in hospital clinics is not directly relevant, and much of the information is implicit, so it is important to simplify this information, remove redundant and useless information, and yet uncover some potentially valuable information. The application of machine learning in healthcare is then essential. Machine learning is a simulation of the data learning process based on data mining, where algorithms are used to build a model and give a judgement result from it. We can therefore process medical data through machine learning methods to obtain a relatively intelligent diagnostic model. This paper focuses on building an early warning model for diabetes based on the RFA. Through tests and experiments, it is found that the prototype system of this DP model can effectively achieve the prediction function of diabetes.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Leila Yousefi, Allan Tucker: Identifying latent variables in Dynamic Bayesian Networks with bootstrapping applied to Type 2 Diabetes complication prediction. Intell. Data Anal. 26(2): 501-524 (2021). https://doi.org/10.3233/IDA-205570

[2] Jafar Abdollahi, Babak Nouri-Moghaddam: Hybrid stacked ensemble combined with genetic algorithms for DP. Iran J. Comput. Sci. 5(3): 205-220 (2021).

[3] Mohammad Zubair Khan, Mangayarkarasi Ramaiah, Vanmathi Chandrasekaran, M. Angulakshmi: Bio-Inspired PSO for Improving Neural Based DP System. J. ICT Stand. 10(2): 179-200 (2021).

[4] Chandrashekhar Azad, Bharat Bhushan, Rohit Sharma, Achyut Shankar, Krishna Kant Singh, Aditya Khamparia: Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus. Multim. Syst. 28(4): 1289-1307 (2021). https://doi.org/10.1007/s00530-021-00817-2

[5] Suja A. Alex, J. Jesu Vedha Nayahi, H. Shine, Vaisshalli Gopirekha: Deep convolutional neural network for diabetes mellitus prediction. Neural Comput. Appl. 34(2): 1319-1327 (2021). https://doi.org/10.1007/s00521-021-06431-7

[6] Nora El-Rashidy, Nesma E. ElSayed, Amir El-Ghamry, Fatma M. Talaat: Prediction of gestational diabetes based on explainable deep learning and fog computing. Soft Comput. 26(21): 11435-11450 (2021).

[7] Simone Faccioli, Andrea Facchinetti, Giovanni Sparacino, Gianluigi Pillonetto, Simone Del Favero: Linear Model Identification for Personalized Prediction and Control in Diabetes. IEEE Trans. Biomed. Eng. 69(2): 558-568 (2021). https://doi.org/10.1109/TBME.2021.3101589

[8] Julian Theis, William L. Galanter, Andrew D. Boyd, Houshang Darabi: Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture. IEEE J. Biomed. Health Informatics 26(1): 388-399 (2021). https://doi.org/10.1109/JBHI.2021.3092969

[9] Hatice Nizam Ozogur, Gokhan Ozogur, Zeynep Orman: Blood glucose level prediction for diabetes based on modified fuzzy time series and particle swarm optimization. Comput. Intell. 37(1): 155-175 (2021). https://doi.org/10.1111/coin.12396

[10] Jobeda Jamal Khanam, Simon Y. Foo: A comparison of machine learning algorithms for DP. ICT Express 7(4): 432-439 (2021). https://doi.org/10.1016/j.icte.2021.02.004

[11] Anand Kumar Srivastava, Yugal Kumar, Pradeep Kumar Singh: Artificial Bee Colony and Deep Neural Network-Based Diagnostic Model for Improving the Prediction Accuracy of Diabetes. Int. J. E Health Medical Commun. 12(2): 32-50 (2021). https://doi.org/10.4018/IJEHMC.2021030102

[12] Asma Ahmed Abokhzam, N. K. Gupta, Dipak Kumar Bose: Efficient diabetes mellitus prediction with grid based RF classifier in association with natural language processing. Int. J. Speech Technol. 24(3): 601-614 (2021). https://doi.org/10.1007/s10772-021-09825-z

[13] P. Preethy Rebecca, S. Allwin: Detection of DR from retinal fundus images using prediction ANN classifier and RG based threshold segmentation for diabetes. J. Ambient Intell. Humaniz. Comput. 12(12): 10733-10740 (2020). https://doi.org/10.1007/s12652-020-02882-3

[14] Shiva Shankar Reddy, Nilambar Sethi, R. Rajender: Rigorous assessment of data mining algorithms in gestational diabetes mellitus prediction. Int. J. Knowl. Based Intell. Eng. Syst. 25(4): 369-383 (2021). https://doi.org/10.3233/KES-210081