International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

An Intelligent Data Mining Model for Diabetic Patients Data Using Hybrid Adaptive Neuro-Fuzzy Inference System

S.Abinesh, G. Prabakaran, R. Arunkumar

Abstract: ***Proper diagnosis of diabetic based on the patient's medical analysis results is an important factor. Data mining helps in analyzing such data includes complex meaningful terms to diagnosis and supports the patients to take remedy action based on the accurate results. The proposed model is a data mining model for analyzing diabetic patient's data using sugeno type adaptive neuro fuzzy inference system with principle component analysis as a hybrid system. The experimental model validated through 200 different data obtained from health clinic with 25 different attributes. The proposed model classifies the data with accuracy of 94.6% where as conventional rough set and k means clustering model produces less classification accuracy of 74.5% and 77.6%.***

Index Terms: ***Diabetic Data, ANFIS, PCA, and Data mining***

I. INTRODUCTION

Fitness is as essential and crucial subject in today's developing world as the technology increases the health issues also increases worldwide. Due to unhealthy life style and lack of physical activities middle aged peoples are affected by chronic diseases. Consumption of fast foods and uneven time management while taking foods leads to diabetics, hypertensions and some cases it leads into cancer etc., Diabetics is a common factor in most of the people now a days and also in few cases children are also affected. Before discussing about data mining models summary about diabetics helps to perform suitable research analysis.

Diabetes mellitus is a fatal disease with rapid growth rate in many countries. Pancreas of the patients suffer and unable to generate necessary insulin to the body. Based on this factor it is classified into three types such as Type1, [8] Type2 and gestational diabetics. In these three types type2 diabetes is considered since the reason for this type2 is mainly due to obesity and improper lifestyle, food habits. Some genetic elements also cause this type2 diabetes for both male and female at any age group. It is very hard particularly for women as it affects the mother and also newborn children in some cases. In the last decade many researches are carried out to develop the early stage detection of diabetics so that prevention can be taken by the patients.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
S. Abinesh*, Research Scholar, Department of Computer Science and Engineering, Annamalai University.
Dr. G. Prabakaran, Associate Professor, Department of Computer Science and Engineering, Annamalai University.
Dr. R. Arunkumar, Associate Professor, Department of Computer Science and Engineering, AnnamalaiUniversity.

Analyzing data for sensing the valuable information changes the trends in bioinformatics and also introduction of statistics and mathematics in analyzing such data changes the dimension of biological data analysis. Research in data mining integrates the process such as data cleaning and data integration which helps in constructing the data warehouses. The data mining process is explained through four major process as data cleaning, data integration, data selection and data transformation. Figure 1.1 gives the illustration of data mining process which involves these four processes.

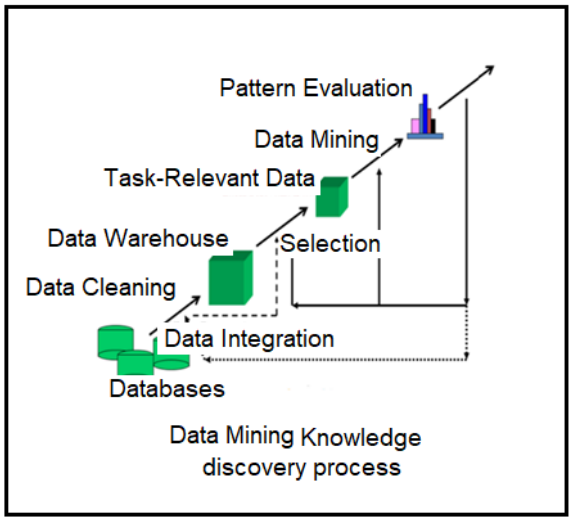


Figure 1.1 Illustration of Data Mining Process

In this data cleaning is used to remove the unwanted noise and other inconsistent data. This inconsistent data leads to false results and affects the performance. While the data integration is used to handle and combine the multiple data sources which is effectively summarizes the data module. Data selection is the third process in data mining where it analysis and retrieves the suitable data from the data base which is relevant for analysis. The last step in data mining is data transformation which converts and consolidates the data into suitable forms. This appropriate forms of data helps in aggregation process. After all this process pattern evaluation and knowledge presentation are considered and these two process are used to identify the necessary patterns and also to represent the data into visualized blocks which is used to provide keen knowledge about the data to the user. In the proposed research model based on the classification issues in conventional models we have tried to improve the classification accuracy of diabetes patient's data using PCA and ANFIS model as a hybrid system

II. RELATED WORKS

A vast survey has been made to summarize the issues in existing research models in diabetes patient data analysis. The research models are used to analyses the advantages of proposed model along with its drawbacks. Literature [1] reports about knowledge discovery database in data mining process. The research model implements data mining process in biomedical data which includes healthcare industry. Data produced in these industries are much higher than others and such data was analyzed in this survey model considering their methods and disciplines. Data mining in web and social networks has discussed in research work [2]. The experimental results indicate the symptoms and treatment starting stages through classification model. MEDLINE data base is used in the proposed model and it converts the data into useful information. While considering the biomedical data, big data also plays a vital role since handling such large data needs special classification and clustering models.

Analysis of big data models have been widely investigated in the literature [3] [4] especially in health care sector where privacy of data is mandatory thereby necessitating encryption based techniques with authentication mechanism integrated into the mining model. Literature [5] reports about a predictive methodology used in big data models to handle the data analysis process efficiently. Hadoop environment is used in the research model for predicting the diabetic types and its treatment methodologies based on results.

Various classification models are already available for data mining process. In case of bio medical data such as cancer data, diabetes data limited algorithms are presented by researchers in this era. A modified version of support vector machine was presented by researcher in literature [6] analyzing the diabetics and hypertension. Similarly a hybrid model is proposed in literature [7] for diagnosing the clinical data of patients. This hybrid model classifies the data using Classification algorithms such as naïve Bayes, decision trees and Bayesian belief networks to obtain results for disease.

Literature [8] also uses support vector machine for non-communicable diseases dataset from world health organization. Analyzing the data set by identifying the treatment types is a unique method followed in the research model. A wide set of experimentation has been carried out under dynamic conditions by considering variation in age groups which is mandatory in mining of diabetes based data.

Literature [9] reports about data mining models used in diabetes data for early prediction. Type 2 diabetes discussed and analyzed in literature [10] which uses similar model which is applied in early stage detection with variation in its parameter selection. So that the results are better than the existing models. Parameters such as insulin level and its secretion level and variation in sugar level are considered to obtain the functional terms in the research model.

Association rule mining (ARM) [11] to improve the data quality in analysis during mining process has been discussed widely in the literature with experimental results. This rules are developed overcome the issues present in conventional models in terms of time and resource intensive. The proposed model rules define the data set into a meaning full one and also consume less time to compute the process. Literature [12] discussed about the instance based algorithm for analyzing large dataset. Based on the aggregating ranks the data and its subsets are related. This research model attains better implementation methodology in terms of data enrich, creative association and patient data structure. Comparing conventional models with proposed adaboost and bagging ensemble model literature [13] provides a new idea towards medical data analysis. The experimental results classifies the data based on their risk factors as well it represents that into decision tree for next step computation.

Literature [14] reports about the issues in identifying the challenges in clinical prediction. A systematic approach was proposed in article which is suitable in data mining applications. This logistical model trains the data set and evaluates based on akaike information criterion and achieves better results. Glycated haemoglobin (HbA1c) and its data set is used in literature [15] and this research helps to analyze the diabetes data also from a large clinical data set. Summarizing a particular data set from a pool of data effectively performed in the research model using cut off values and novel biomarkers. The classification accuracy of this model is about 85.63% and the performance of the system need to be increased for further improvement in classification accuracy.

From the survey about various data models it is observed that the issues present in classifying the data in case of large clinical data pool. So an effective algorithm is needed to reduce the large data set so that analysis process will be ease for better improvement. To resolve all the issues we are proposing a hybrid principle component based adaptive neuro fuzzy inference model to handle large clinical data set effectively. Section III provides the mathematical model of proposed work and section IV discusses the experimental results.

III. PROPOSED WORK

The mathematical model of proposed data mining model is presented in this section by summarizing the ANFIS model and principal component analysis model. Since PCA is a well-known model used to reduce the data size effectively. Many big data models proposed PCA as essential model in reducing the data size. Adaptive neuro fuzzy inference system basically used fuzzy and neural network model for classification. Based on logical system and nonlinear mapping of inputs the fuzzy domain operations are performed. Fuzzy inference engine and its rules converts the inputs into fuzzy sets so that the results are transformed into arithmetical values. Gaussian function plays an important role in fuzzy sets and linear functions are used to define the output rules in ANFIS model along with standard deviation and mean. The membership function and its coefficients are used to define the network parameters. In the proposed model sugeno fuzzy model is used as if-then rules and typical rule is given as

$$If\ x\ is\ A\ and\ y\ is\ B\ then\ x = f(x; y) \qquad (2.1)$$

The rule given above is anterior and the resulting function is obtained as $z = f(x, y)$ and this mostly described as polynomial function. In case of exponential model the function varies along with output of the system. In first order sugeno model along with first order polynomial the fuzzy model and FIS constants are considered as a special case inference system. Reducing the first order Sugeno fuzzy inference system having two rules.

The proposed model design consists of two modules as antecedent and conclusion phase. This is due to reduce the computational complexity in defuzzification. Figure 2.1 gives an illustration of proposed ANFIS model with two inputs, five layer and single output.

$$Rule\ 1: If\ X\ is\ A_1\ and\ Y\ is\ B_1;\ then\ f_1 = p_1x + q_1y + r_1 \quad (2.2)$$

$$Rule\ 2: If\ X\ is\ A_2\ and\ Y\ is\ B_2;\ then\ f_2 = p_2x + q_2y + r_1 \quad (2.3)$$

The proposed model design consists of two modules as antecedent and conclusion phase. This is due to reduce the computational complexity in defuzzification. Figure 2.1 gives an illustration of proposed ANFIS model with two inputs, five layer and single output. The two phases used in the model are linked through networks and each network has several nodes. The proposed model reflects a multi layer neural network model with two inputs and a single decision based output. A five layer structure has been utilized in the proposed model where inputs are connected to the second layer which is activated by a set of membership functions which define the fuzzy rules. At this stage, the input values are transformed into fuzzy values and the rules based on membership functions are applied in the next layer following which it is normalized in the succeeding layer by obtaining inputs from all nodes or neurons. The final layer represents the reverse process which provides the decision based on the firing strength of the fuzzy inputs.

The merits of ANFIS over conventional fuzzy or neural models are that they adapt themselves to changing input values and decisions are made based on the learning process of the neurons. This learning process is activated by a set of weights which update depending upon the convergence criteria and iterations are stopped or terminated when the weight update gets more or less saturated indicating optimal convergence towards the ideal solution.

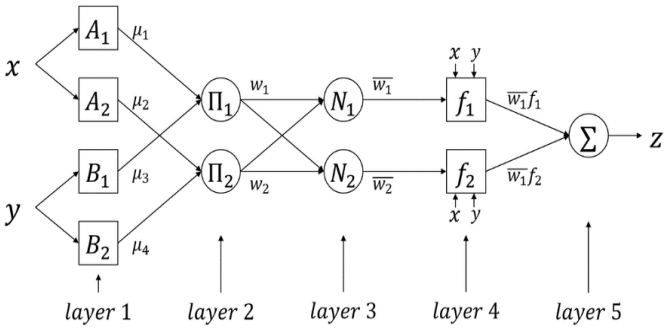


Figure 2.1 ANFIS model

The layer 1 performs execution and also provides membership function to each input based on the Gaussian functions and this is summarized into

$$l_1^{\,2} = \mu_a = e^{\left(\frac{-(x-h)^2}{\delta^2}\right)} \quad (2.4)$$

The layer 2 executes the fuzzy AND present in the results of layer 1based on fuzzy rules and this is given as

$$l_2^{\,2} = w_a = \mu_a(x_1) \times \mu_b(x_2) \quad (2.5)$$

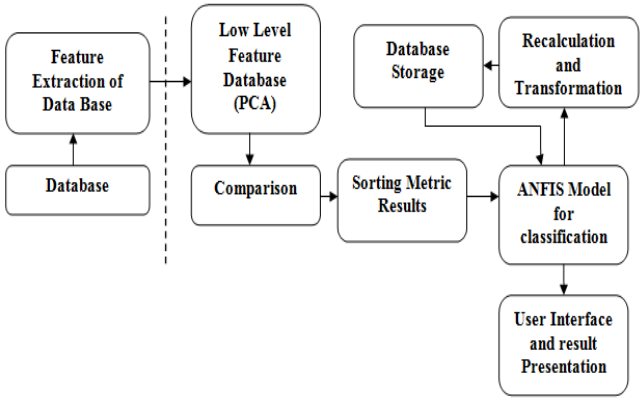
The values obtained from layer 2 are need to be normalized and layer 3 performs such normalization process

$$l_3^{\,2} = \overline{w_a} = \frac{w_a}{\sum_i^4 = w_a} for\ i = 1,2,3,4 \quad (2.6)$$

Layer 4 and layer 5 executes the fuzzy rules and defuzzification process and it is given as

$$l_4^{\,2} = \overline{w_a y_a} = \overline{w_a}\left(\alpha_1^i x_1 + \alpha_2^i x_2 + \alpha_3^i\right) i = 1,2,3,4 \quad (2.7)$$

In this hybrid learning model training the algorithm performed by forward pass and backward pass. The forward pass considers all the parameters before initializing the functional signals up to fourth layer and LSE was identified for the consecutive values. If the parameters are identified functional signal goes into forward until it calculates the error. While in backward process the error rates are considered backwards and all the parameters are rationally presented to obtain the function.



Figur2e 2.2 Illustrative Description of Proposed Model

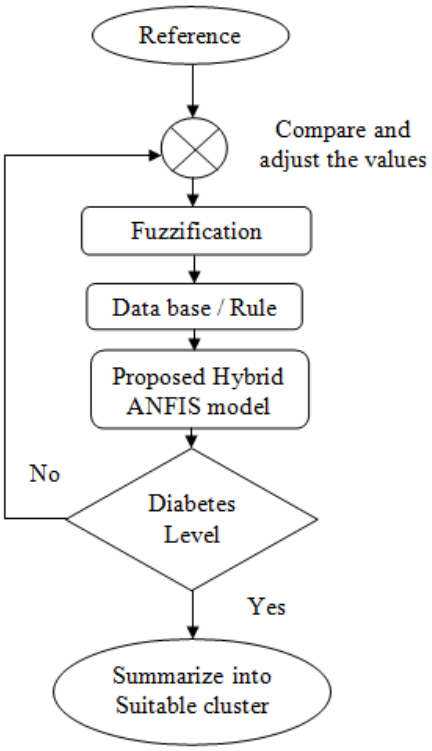


Figure 2.3 Process Flow of Proposed Hybrid ANFIS Model.

Figure 2.2 gives the proposed model process as an illustrative description. The process starts from data base feature extraction and these features are large in size while the PCA is used to obtain the low level featured data base.

So the comparison can be performed based on the threshold levels. Then the sorted results are given to ANFIS model which performs classification of sorted data by recalculating and transformation process. The final results are provided to the user through user interface modules. The process flow is given in figure 2.3 and it starts from reference values and the values from the controlled output. Both the values is compared and then it is fuzzification is performed followed by rule based and data based functions then it is applied to artificial neural network using back propagation algorithm and this section is considered as Hybrid ANFIS model and finally it checks the values and maintain the dataset for the patient automatically.

The root mean square error the functional model is calculated through the given equation

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(d_i - l_i)^2\right)} \quad (2.8)$$

Where d_i is the desired output and l_i is the ANFIS output for the sample from training data and n is the number of training samples.

IV. RESULT AND DISCUSSION

The proposed Hybrid model is tested on a 1.8 GHz processor with 2GB RAM running Windows 7 and coded by Mat lab 14.1. The AIM'94Dataset is taken in account for analyzing the proposed model. 73 patient's data set has more than 700 features in that has 5 features are taken in account for simulations and the process is carried out with other two models and proposed hybrid ANFIS model. The parameters used for analysis were the time taken to complete, accuracy of results and the scalability of the algorithms with respect to the size of the data.

Data was passed to the algorithms as random feature values and records. Since data reduction and prediction of diabetes is essential and the necessary conditions are not depends upon the serial values. So a random data approach was considered in order to find the better optimization result. Table 3.1 depicts the list of feature selected for processing using PCA. High dimensional diabetes data was reduced into low dimensional data using PCA, in this case the data set contains of 700 features and out of this a random of 100 data feature was selected based on the principal component analysis.

Dataset	**Selected Features from original dataset** (75 Features)
AIM'94 Dataset (73*700 Features)	1735, 2212, 4470, 4505, 5835, 5947, 5966, 6143, 6185, 6327, 6512, 6529, 6541, 6566, 6836, 7005, 7211, 7459, 7782, 7910, 8109, 8662, 8667, 8979, 9894, 10075, 10185, 3767, 14212, 14810, 14816, 15896, 15898, 16004, 16559, 17337, 17566, 18321,18811, 18974, 19044, 19646, 20043, 20173, 20437,

Table 3.1 Features Selected by PCA used in the Proposed Model

The membership function of proposed model is illustrated in figure 3.1. It is observed that the membership function resembles same as Gaussian function so the proposed model behaves similar to Gaussian function.

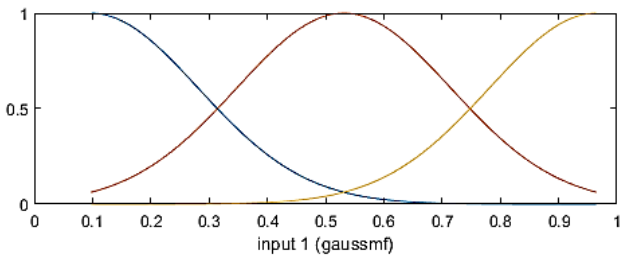


Figure 3.1 Membership Function of Proposed Model using Gaussian distribution

The results are analyzed in terms of true positive, true negative, false positive and false negative. Table 3.2 gives summary about the condition used in proposed model analysis.

True positive	Input is determined as healthy with optic nerve diagnosed
True Negative	Input is determined as healthy
False Positive	Input is determined as a patient
False Negative	Input is determined as healthy with optic nerve diagnosed

Table 3.2 Conditions used in Experimental Model

The diagnosing accuracy was obtained from the classification results of proposed hybrid ANFIS model and it is given as

$$Diag_{accu} = \frac{\sum_1^k a(h_k)}{|h|} \quad (9)$$

Figure 3.2 shows the error performance n for the epochs based on the data set which is given as input. The error performance is calculated based on the no of errors occurs.

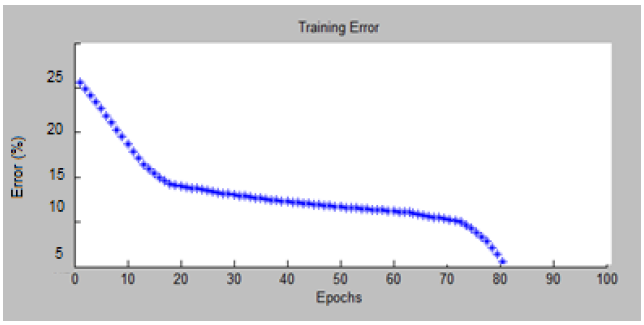


Figure 3.2 Error Performance of Proposed Model after Training

It is observed that the error percentage initially high and the no of epoch's increases the error percentage gradually decreases in the proposed hybrid ANFIS model. Figure 3.3 depicts the comparison between the training data and the FIS output and the blue dots mentions the training data and red star indicates the FIS output.

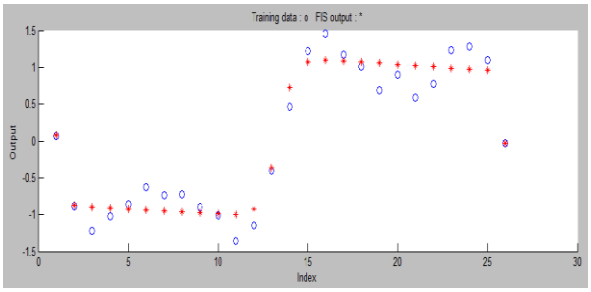


Figure 3.3 Training Data and FIS Output

The diagnosis of diabetes results is based on sensitivity and specificity analysis and it is given as

$$sensitivity = \frac{TP}{TP + FN} (\%) \quad (10)$$

$$specificity = \frac{TN}{FP + TN} (\%) \quad (11)$$

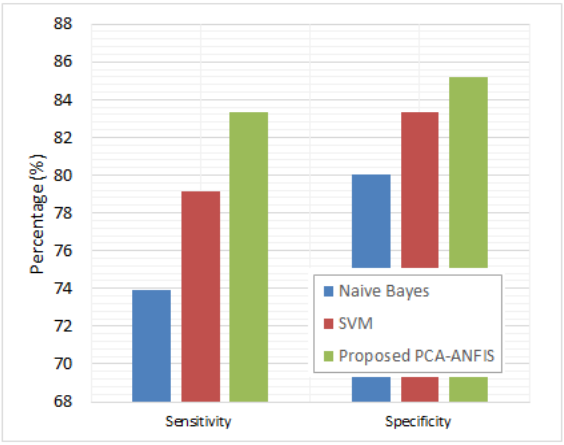


Figure 3.4 Sensitivity and Specificity Comparison of Proposed Model

Figure 3.4 gives the comparison plot of sensitivity and specificity of the proposed model. It is observed that the proposed model attains better sensitivity and specificity than conventional models such as k-Means, Roughs set.

Sl..No.	Method	Accuracy (%)
1	k-Means	74.5
2	Rough set	77.6
3	Proposed PCA-ANFIS	94.6

Table 3.3 The Classification Accuracies of this PCA-ANFIS System and Previous Methods.

Table 3.3 shown above presents the overall classification accuracy of the proposed ANFIS model and its performance compared against benchmark techniques such as rough set and K-means cluster based approach. A significant improvement of nearly 24 % improvement in the classification accuracy is registered which justifies the superiority of the proposed supervised learning model. Considering these two factors the classification accuracy of proposed model reaches about 94.61% which is much better than conventional modes.

V. CONCLUSION

Utilization of data mining techniques have been on the rise especially with the advent of big data and related concepts. Data mining techniques have been effectively used in medical sectors of health care for retrieval of records from archives, prediction of occurrence of certain medical conditions based on the symptoms and age related data where mining proves to be an effective methodology. A novel supervised and hybrid adaptive neuro fuzzy inference model has been proposed for detection of diabetes based on an extensive data set related to various age groups monitored and recorded over a considerable period of time.

The proposed research model uses hybrid Adaptive Neuro Fuzzy Inference System for analyzing the diabetic patient's data. Principal component analysis has been integrated in the proposed work to reduce the dimensionality of the feature vectors thus accounting the reduced time consumption and computational overhead. Proposed model performs better than conventional models such as k means and rough set models. The classification accuracy of the proposed model attains 94.61% which is 24% greater than the conventional models. The future scope of this research would be an optimization model in order to increase the efficiency by considering various diagnostic methods and different feature extraction, classifier methods. Learning based models are best suited for mining based applications where manual intervention is maximally eliminated since the network is able to predict the occurrence of a particular condition based on the training imparted to it.

REFERENCES

1. NeeshaJothi, Nur'Aini Abdul Rashid, Wahidah Husain, "Data Mining in Healthcare – A Review" Procedia Computer Science, Volume 72, Pages 306-313, 2015
2. FarhiMarir, Huwida Said, Feras Al-Obeidat, "Mining the Web and Literature to Discover New Knowledge about Diabetes" Procedia Computer Science, Volume 83, Pages 1256-1261, 2016
3. Prableen Kaur, Manik Sharma, Mamta Mittal, "Big Data and Machine Learning Based Secure Healthcare Framework" Procedia Computer Science, Volume 132, Pages 1049-1059, 2018
4. Nelson Sanchez-Pinto.L, Yuan Luo, Matthew M. Churpek, "Big Data and Data Science in Critical Care" Chest, Volume 154, Issue 5, Pages 1239-1248, 2018
5. Saravanakumar. N. M, Eswari.T, Sampath.P, Lavanya.S, "Predictive Methodology for Diabetic Data Analysis in Big Data" Procedia Computer Science, Volume 50, Pages 203-208, 2015
6. Namrata Singh, Pradeep Singh, DeepikaBhagat, "A rule extraction approach from support vector machines for diagnosing hypertension among diabetics" Expert Systems with Applications, Volume 130, Pages 188-205, 2019
7. Leonard Barreto Moreira, Anderson AmendoeiraNamen, "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia" Computer Methods and Programs in Biomedicine, Volume 165, Pages 139-149, 2018
8. Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients" Journal of King Saud University - Computer and Information Sciences, Volume 25, Issue 2, Pages 127-136, 2013
9. Mohammed Zeyad Al Yousef, MazenFerwana, SherifSakr, Riyad Al Shammari, "Early prediction of diabetes by applying data mining techniques" Computer Methods and Programs in Biomedicine, Volume 171, Page 3, 2019
10. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining" Informatics in Medicine Unlocked, Volume 10, Pages 100-107, 2018

11. MingkaiPeng, VijayaSundararajan, Tyler Williamson, Evan P. Minty, HudeQuan, "Exploration of association rule mining for coding consistency and completeness assessment in inpatient administrative health data" Journal of Biomedical Informatics, Volume 79, Pages 41-47, 2018
12. JorisFalip, Amine Aït-Younes, Frédéric Blanchard, Brigitte Delemer, Michel Herbin, "Visual instance-based recommendation system for medical data mining" Procedia Computer Science, Volume 112, Pages 1747-1754, 2017
13. SajidaPerveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes" Procedia Computer Science, Volume 82, Pages 115-121, 2016
14. FaridehBagherzadeh-Khiabani, AzraRamezankhani, FereidounAzizi, FarzadHadaegh, DavoodKhalili, "A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results" Journal of Clinical Epidemiology, Volume 71, Pages 76-85, 2016
15. Herbert F. Jelinek, Andrew Stranieri, Andrew Yatsko, SitalakshmiVenkatraman, "Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis" Computers in Biology and Medicine, Volume 75, Pages 90-97, 2016