Type II Diabetes Mellitus Analysis Based On Linear Regression and Cascade Forward Neural Network (CFNN)

Rajeev Rajput[1], Dr Jitendra Agrawal[2], Prof . Sumit Sharma[3]
M.Tech Scholar CSE[1], Head of Department CSE[3] , Professor IT[2]
Vashnavi Institute of Technology (VIT)[13,], School of Information Technology RGPV[2]
Bhopal [M.P.]

Abstract : In this research work present a hybrid model that is combination of linear regression model and cascaded feed forward network with back propagation training. The detection of the diabetic type 2 patient shows better accuracy as compare to other previous methods. In the different previous method analysis got the drawbacks of previous method that is over come in the proposed method that is shown in the result discussion. The outcome proposed method shown that can be detected with an awfully high accuracy, up to 96.54 % in the different diabetic patient datasets. The proposed method also shows lower complexity and better timing due to regression model soft threshold values. In the proposed hybrid model is the combination of CFFNN and Linear Regression method. Cascade-forward networks are similar to feed-forward networks, but include a connection from the input and every previous layer to following layers. As with feed-forward networks, a two-or more layer cascade network can learn any finite input-output relationship arbitrarily well given enough hidden neurons.For greater accuracy on low-dimensional through medium-dimensional data sets, fit a linear regression model.

Index Terms - Diabetic Type 2, Cascaded Feed Forward Network , Linear Regression, Machine Learning Processes, Random Forent And Accuracy.

I. =INTRODUCTION

Diabetes mellitus could be a heterogeneous group of diseases specify by chronic elevation of glucose within the blood. It arises as a result of the body is unable to provide enough insulin for its own desires, either due to impaired endocrine secretion, impaired endocrine action, or both. polygenic disorder affects some three hundred million folks world-wide, and is on the rise. Chronic exposure to high glucose may be a leading explanation for kidney disease, visual loss and a variety of different kinds of tissue injury. sickness} additionally predisposes to blood vessel disease, not least as a result of it's typically in the middle of cardiovascular disease, supermolecule disorders and blubber. several cases of polygenic disease and most of its unwanted semipermanent consequences ar probably avertible, however this may need intervention at a social moreover as at a medical level. This section of Diapedia offers associate introduction to the history of polygenic disorder, its clinical presentation, its current classification and its international medicine. we tend to conjointly introduce a number of the psychological and social aspects of polygenic disease, together with the 'hot topics' that dominate the media, and supply an summary of current areas of analysis interest. of these topics ar thought-about in bigger detail elsewhere in Diapedia, and that we hope you'll explore them additional.

Diabetes Mellitus ("diabetes" for short) may be a serious illness that happens once your body has issue properly regulation the number of dissolved sugar (glucose) in your blood stream. it's unrelated to a equally named disorder "Diabetes Insipidus" that involves kidney-related fluid retention issues. so as to grasp polygenic disease, it's necessary to initial perceive the role aldohexose plays with reference to the body, and what will happen once regulation of aldohexose fails and glucose levels become perilously low or high. The tissues and cells that structure the anatomy reside things, and need food to remain alive. The food cells eat could be a kind of sugar referred to as aldohexose. mounted in situ as they're, the body's cells ar utterly captivated with the blood stream during which they're bathed to bring aldohexose to them. while not access to adequate aldohexose, the body's cells don't have anything to fuel themselves with and shortly die. citizenry eat food, not aldohexose.

Human foods get regenerate into aldohexose as a vicinity of the conventional digestion method. Once reborn, aldohexose enters the blood stream, inflicting the extent of dissolved aldohexose within the blood to rise. The blood stream then carries the dissolved aldohexose to the assortedtissues and cells of the body. tho' aldohexose could also be out there within the blood, close cells aren't ready to access that aldohexose while not the help of a chemical internal secretion referred to as endocrine. internal secretion acts as a key to open the cells, permitting them to receive and utilize on the market aldohexose. Cells absorb aldohexose from the blood within the presence of internal secretion, and blood glucose levels drop as sugar leaves the blood and enters the cells. internal secretion will be thought of as a bridge for aldohexose between the blood stream and cells. it's vital to grasp once levels of endocrine increase, levels of sugar within the blood decrease (because the sugar goes into the cells to be used for energy).The body is meant to control and buffer the number of aldohexose dissolved within the blood to take care of a gradual provide to satisfy cell wants.

releases endocrine into the blood stream to bring aldohexose levels backpedal. The concentration of aldohexose accessible within the blood stream at any given moment depends on the number and sort of foods that individuals eat. Refined carbohydrates, candy and sweets ar straightforward to interrupt down into aldohexose. Correspondingly, glucose levels rise speedily when such foods are consumed. In distinction, blood sugars rises step by step and slowly when consumption a lot of advanced, unrefined carbohydrates (oatmeal, apples, baked potatoes, etc.) that need a lot of biological process steps crop up before aldohexose will be yielded. baby-faced with speedily rising glucose concentrations, the body should react quickly by emotional giant amounts of endocrine all directly or risk a dangerous condition referred to as Hyperglycemias (high blood sugar) which is able to be represented below. The inflow of endocrine permits cells to utilize aldohexose, and aldohexose concentrations drop. whereas

aldohexose levels will rise and fall quickly, endocrine levels modification rather more slowly.The main objective of proposed research work to analysis the dubieties type 2 patient and give the prediction of possibility of deities. For this first requirement is the analysis of data set and its different attributes in it. Further improve the accuracy of proposed work with the help of ANN based machine learning method.

II LITERATURE SURVEY

Wu, Kopitar, Leon, et al. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models", In this reseach work authors present results that using new data in the EHR system to rebuild prediction models not only improves prediction performance, but also stability of the variable importance ranking, although not equally in diferent machine learning prediction models. Our results found no clinically relevant improvement when employing machine learning-based models over the more conventional regression models in terms of predictive performance. Even with calibration of the models, visualization of the observed versus actual FPGL showed some advantages in using simpler models. When observing the stability of variable ranking based on relative importance of variables, one can notice that a method like LightGBM results in much more stable results in comparison to other methods, which were more prone to high variance in variable importance. Both regression-based methods also proved as comparable alternatives. Since regression-based prediction models have been regularly used in clinical practice they could represent a better alternative in some clinical environments. Te results in this study show signifcant improvement in terms of AUC, AUPRC and RMSE for all tested methods as the amount of collected data increases. For all tested predictive models in most of the experiments, we were able to show that additional data availability positively correlates with improved predictive performance and more stable variable importance-based ranking of variables. Te opportunity of updating models arises as additional routine data become available over time. Future research needs to explore the implementation of different approaches of building ensemble methods. In this case, stacking and blending of different prediction models could be taken into consideration. However, such systems bring along even more challenges in terms of interpreting the results that should support decisions of the healthcare experts[1].

Krishnan, Devi R., et al. "Evaluation of predisposing factors of Diabetes Mellitus post Gestational Diabetes Mellitus using Machine Learning Techniques" , In this study, we tested various ML classification models on 15 feature dataset of 77 patients. In clinical practice, predicting the risk of DM after GDM is a challenging task. In our research, we make use of prognostic machine learning(ML) techniques that inputs 15 predisposing factors and outputs the risk of DM. The results from our study demonstrated that Random Forest and Gaussian Naive Bayes classifies the data comparatively better than the other models[2].

Vigneswari, et.al. "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus", Machine learning tree classifiers were analyzed for predicting diabetes mellitus in this study. The accuracy of these classifies were calculated based on the True Positive, True Negative, False Positive, and False Negative of the classifiers. In this study, Logistic Model Tree (LMT) classifier achieved an accuracy of 79.31% with an average time of 0.49 sec to build the model which is higher than the Random Forest tree classifier with 78.54% accuracy and 0.04 sec. Machine learning tree classifiers other than the ones considered here can be analyzed in future for constructing hybrid tree classifier algorithms [3].

Alassaf, Reem A. et.al., ""Preemptive Diagnosis of Diabetes Mellitus Using Machine Learning" In this paper, Saudi Arabia medical data has been explored for the first time in diagnosing diabetes mellitus. Correlation coefficient and recursive feature elimination were used for feature selection, then four classification algorithms, namely: ANN, SVM, NB, and K-NN were compared on the basis of classification accuracy, precision, recall and f-measure. ANN outperformed the rest with classification accuracy of 77.5%. Potential future work could be in improving the classification accuracy by investigating different feature selection methods or classifiers. Moreover, decreasing the FN rate in mis-classification, has high significance in the context of medicine [4].

Ioanniskava Kiotiset "Machine Learning and Data Mining Methods in Diabetes Research", [2017]The growth of biotechnology and health sciences have led to a significant production of data, such as high throughput genetic data and clinical information, generated from large Electronic Health Records (EHRs). To this end, application of machine learning and data mining methods in biosciences is presently, more than ever before, vital and indispensable in efforts to transform intelligently all available information into valuable knowledge. Diabetes mellitus (DM) is defined as a group of metabolic disorders exerting significant pressure on human health worldwide. Extensive research in all aspects of diabetes (diagnosis, etiopathophysiology, therapy, etc.) has led to the generation of huge amounts of data. The aim of the present study is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the first category appearing to be the most popular[5].

PhattharatSongthung et.al, "Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification", [2016]Diabetes is a chronic disease that contributes to a significant portion of the healthcare expenditure for a nation as individuals with diabetes need continuous medical care. In order to prevent or delay the onset of type 2 diabetes, it is necessary to identify high risk populations and introduce behaviour modifications as early as possible. Screening the population to identify high risk individuals is an important task. One of the most accurate tests of diabetes is through the analysis of fasting blood sugar, but it is invasive and costly. Furthermore, it is only useful when the individual is already displaying symptoms i.e., making a diagnosis, which is considered too late to be an effective screening mechanism [6].

Madhuri Panwar et.al."K-nearest neighbour basedmethod-ology for accurate diagnosis of diabetes mellitus", Diabetes is one of the leading causes of death, disability and economic loss throughout the world. Type 2 diabetes is more common (90-95% worldwide) type of diabetes. However, it can be prevented or delayed by taking the right care and interventions which indeed an early diagnosis. There has been much advancement in the field of various machine learning algorithms specifically for medical diagnosis. But due to partially complete medical data sets, accuracy often decreases, results in more number of mis-classification that can lead t o harmful complications. An accurate prediction and diagnosis of a disease becomes a challenging research problem for many researchers. Therefore, aimed to improve the diagnosis accuracy we have proposed a new methodology, based on novel pre-processing techniques, and K-nearest neighbour classifier**[7].**

III. Proposed Method

In this chapter discuss the proposed solution, in which solve the problem of previous work that is Shown in previous chapter -2 literature survey. In the above chapter discuss the proposed method for Type 2 diabetes mellitus prediction. For the detection of the detection as well as predication of the use hybrid model that is based on neural network and regression.

Diabetes Patient Prediction Model

In this section, discuss the diabetes patient prediction architecture and define the detection process model. In general diabetes patient prediction first train the system with the help of data set, in the data set there are different body such as Age, BMI, Impaired glucose tolerance (IGT), Family history of DM, Lower high-density lipoprotein cholesterol or (HTG), Hypertension or cardiovascular disease. All these parameters are used to train proposed system.

Linear Regression In statistics, regression may be a linear approach to modeling the link between a variable quantity and one or independent variables.

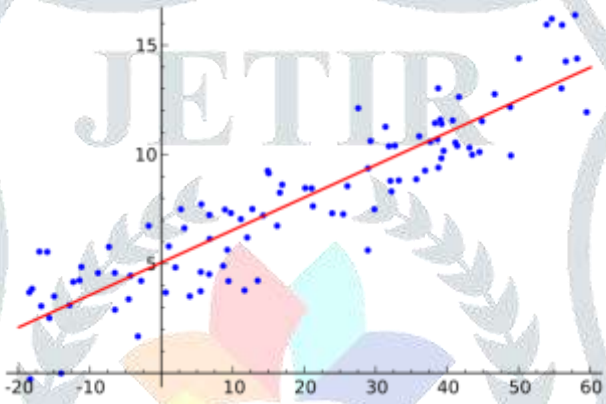

Fig. 2 Shows Linear Regression

Given a date set $\{y_i, x_{il}, \dots x_{ip}\}\ \frac{n}{i=1}$of n statistical unites, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regresses **x** is linear.

$$yi = \beta_0 1 + \beta_1 x_{i\,1} + \cdots + \beta p_{x\,i_p} + \epsilon_i \quad = x_i^T\beta + \varepsilon_{i,} i = 1, \dots, n, \qquad 1$$

where ^{T} denotes the transpose. so that $x_i^T\beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in matrix notation as.

$$y = x\beta + \varepsilon, \qquad 2$$

Where

$$Y = \begin{pmatrix} y1 \\ y2 \\ \vdots \\ yn \end{pmatrix} \qquad 3,$$

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{matrix} 1 & x_{11} \dots & x_{1p} \\ 1 & x_{21} \dots & x_{2p} \\ & & \vdots \\ 1 & x_{n1} \dots & x_{np} \end{matrix} \qquad 4$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_P \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad 5.$$

Cascade-Forward Neural Network

In perception connection that is formed between input and output is a form of direct relation while in FFNN connection formed between input and output is indirect relationship. The connection is nonlinear in shape through an activation function in the hidden layer. If the connection form on perception and multilayer network is combined, then the network with direct connection between the input layer and the output layer is formed, besides the connection indirectly. The network formed from this connection pattern is called Cascade Forward Neural Network (CFNN). The equations are formed from the CFNN model can be written as follows:

$$\sum_{i=1}^{n} f^{i}\omega_{t}^{i}\chi_i + f^{0}\left(\sum_{j=1}^{n} \omega_j^0 f_j^h\left(\sum_{i=1}^{n} \omega_{ji}^h \chi_i\right)\right)$$

Where $f^i$is the activation function from the input layer to the output layer and ω_{ji}^h is weight from the input layer to the output layer. If a bias is added to the input layer and the activation function of each neuron in the hidden layer is then equation (4.6) become

$$y = \sum_{i=1}^{n} f^i\,\omega_i^i\chi_i + f^0\left(\omega^b + \sum_{j=1}^{k} \omega_j^0 f^h\left(\omega_j^b + \sum_{i=1}^{n} \omega_{ji}^h \chi_i\right)\right) \qquad 4.6$$

In this research, the CFNN model is applied in time series data. Thereby, the neurons in the input layer are the lags of time series data $\mathrm{Xt}-1, \mathrm{Xt}-2, \ldots, \mathrm{Xt}-\mathrm{p}$, whereas the output is the current data Xt . The architecture of CFNN model in predicting time series is shown at Fig. 4.3.

IV.Simulation And Result

In this chapter discuss the simulation model and result of proposed algorithm. For the implementation of proposed algorithm use Matrix laboratory. Matrix laboratory is a well-known tool for such kind of algorithm implementation related to data analysis calculation. MATLAB contain a rich function family of data analysis The result of proposed method for development of middle ware using machine learning technique for diabetic type 2 shown in this section, simulation of our proposed method and result calculation. We have done proposed work with the help the MATLAB R 2015a (8.1.0.602) software and simulate our whole proposed methodology in data analysis. Basic configuration of our system is: Processor: Intel (R) Quad Core (VM) i3 – 3110 Central Processing unit @, 2.40 GHz with 4GB RAM: System type: 64-bit Operating System.

Result Comparison

In the above section discuss the result compression of proposed method. Now compare the resultant output with different machine learning methods that is shown in below. In the below result table 5.1 shows the accuracy of classification of different methods using machine learning approach. The resultant accuracy of proposed work is near around 79% (without regression).

Table 4.1 The values of accuracy of classification made on pima Indian diabetes dataset

Method	Accuracy(%)
Proposed Method	79.6
Discrim	77.5
MLP	73.8
Logdisc	78.2

the accuracy of the comparison of different methods such as Discrim, MLP, Logdisc, SMART, Bayesnet, NaivebAY, Random Forent, J48, SGD, SMO, Backiprop, RBF, LMT , in the graphical In all the methods apply direct classification techniques on different machine learning platform such as Weka, R , R studio and Python. Our proposed Cascaded method shows better accuracy near around 80%.

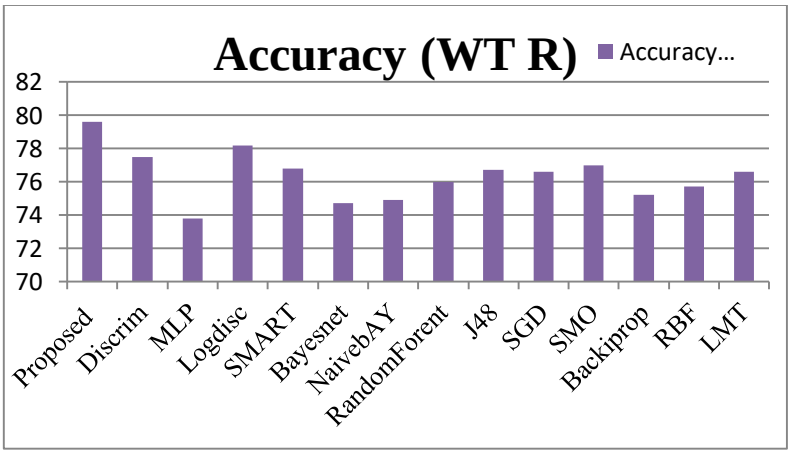

Fig. 3 Shows the accuracy comparison with different methods

In the table 4.2 shows the result comparison of proposed method with regression in all approaches. In the references [1] researchers use logistic regression with K mean clustering approach.

Table 2 Result Comparison

Method	Accuracy	Reference
Proposed	**96.54%**	**2020**
Base Paper	95.42%	Wu, Han, [1]
HPM	92.38%	B.M. Patil [6]
AMMLP	89.42%	Alexis Marcano-Cedeno [8]
J48(pruned)	89.93%	Aliza Ahmad [7]
J48(unpruned)	89.35%	Aliza Ahmad [7]

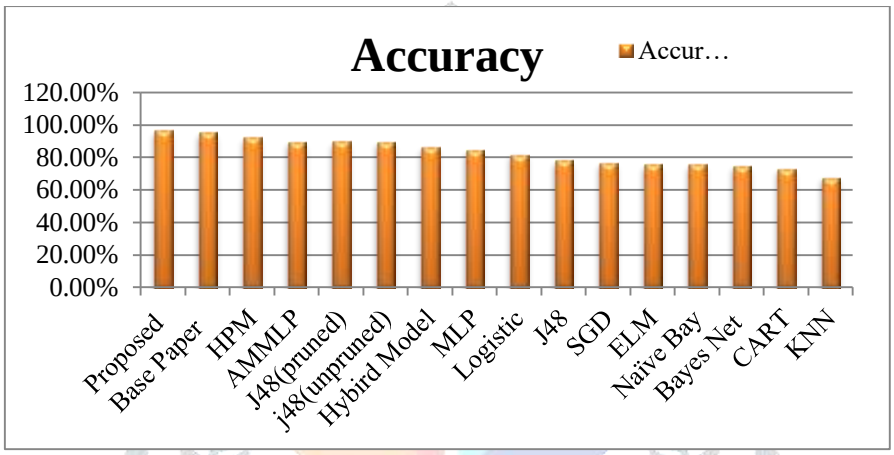

Fig. 4 Shows the accuracy comparison with different methods (With Regression)

As clearly shows that proposed method shows better result as compared to other previous methods. The proposed method shows better result due to the hybrid model of three different methods first one is data cleaning and data correction using linear regression method, after that apply back propagation based learning method and classify by threshold based cascaded neural network. .

V. Conclusion

In the research work present a hybrid model that is combination of linear regression model and cascaded feed forward network with back propagation training. The detection of the diabetic type 2 patient shows better accuracy as compare to other previous methods. The outcome proposed method shown that can be detected with an awfully high accuracy, up to 96.54 % in the different diabetic patient datasets. The proposed method also shows lower complexity and better timing due to regression model soft threshold values. The result comparison of proposed method shown in the different tabular form in the table 1 and 2.

References

[1] Wu, Kopitar, Leon, Primoz Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." Scientific reports 10, no. 1 (2020): 1-12.

[2] Krishnan, Devi R., Gayathri P. Menakath, Anagha Radhakrishnan, Yarrangangu Himavarshini, A. Aparna, Kaveri Mukundan, Rahul Krishnan Pathinarupothi, Bithin Alangot, Sirisha Mahankali, and Chakravarthy Maddipati. "Evaluation of predisposing factors of Diabetes Mellitus post Gestational Diabetes Mellitus using Machine Learning Techniques." In 2019 IEEE Student Conference on Research and Development (SCOReD), pp. 81-85. IEEE, 2019..

[3] D. Vigneswari1 , N. Komal Kumar , V. Ganesh Raj , A. Gugan , S. R.Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus", 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019.

[4] Alassaf, Reem A., Khawla A. Alsulaim, Noura Y. Alroomi, Nouf S. Alsharif, Mishael F. Aljubeir, Sunday O. Olatunji, Alaa Y. Alahmadi, Mohammed Imran, Rahma A. Alzahrani, and Nora S. Alturayeif. "Preemptive Diagnosis of Diabetes Mellitus Using Machine Learning." In 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1-5. IEEE, 2018.

[5] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).

[6] Songthung, Phattharat, and KunwadeeSripanidkulchai. "Improving type 2 diabetes mellitus risk prediction using classification." In Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on, pp. 1-6. IEEE, 2016.

[7] Panwar, Madhuri, Amit Acharyya, Rishad A. Shafik, and Dwaipayan Biswas. "K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus." In Embedded Computing and System Design (ISED), 2016 Sixth International Symposium on, pp. 132-136. IEEE, 2016.

[8] Saxena, Kanak, and Richa Sharma. "Diabetes mellitus prediction system evaluation using c4. 5 rules and partial tree." In Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2015 4th International Conference on, pp. 1-6. IEEE, 2015.

[9] ZahirniaKiarash, Mehdi Teimouri, RohallahRahmani, and Amin Salaq. "Diagnosis of type 2 diabetes using cost-sensitive learning." In Computer and Knowledge Engineering (ICCKE), 2015 5th International Conference on, pp. 158-163. IEEE, 2015.

[10] Lee, Bum Ju, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim. "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes." IEEE journal of biomedical and health informatics 18, no. 2 (2014).

[11] Choi, SooBeom, Won Jae Kim, Tae KeunYoo, JeeSoo Park, Jai Won Chung, Yong-ho Lee, EunSeok Kang, and Deok Won Kim. "Screening for prediabetes using machine learning models." Computational and mathematical methods in medicine 2014 (2014).

[12] 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.

[13] Dalakleidi, Kalliopi V., KonstantiaZarkogianni, Vassilios G. Karamanos, Anastasia C. Thanopoulou, and Konstantina S. Nikita. "A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in Type 2 Diabetes patients." In Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on, pp. 1-4. IEEE, 2013.

[14] OlokobaAbdulfatai B., Olusegun A. Obateru, and Lateefat B. Olokoba. "Type 2 diabetes mellitus: a review of current trends." Oman medical journal 27, no. 4 (2012).

[15] Marinov, Miroslav, Abu Saleh Mohammad Mosa, IllhoiYoo, and Suzanne Austin Boren. "Data-mining technologies for diabetes: a systematic review." Journal of diabetes science and technology 5, no. 6 (2011).

[16] Sacks, David B., Mark Arnold, George L. Bakris, David E. Bruns, Andrea Rita Horvath, M. Sue Kirkman, AkeLernmark, Boyd E. Metzger, and David M. Nathan. "Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus." Clinical chemistry 57, (2011).

[17] Zhang, Ping, Xinzhi Zhang, Jonathan Brown, DorteVistisen, Richard Sicree, Jonathan Shaw, and Gregory Nichols. "Global healthcare expenditure on diabetes for 2010 and 2030." Diabetes research and clinical practice 87, no. 3 (2010): 293-301.

[18] B. M. Patil, R. C. Joshi and DurgaToshniwal, "Association rule for classification of type -2 diabetic patients", Second International Conference on Machine Learning and Computing, 2010.

[19] Rodbard, Helena W., Andrew J. Green, Kathleen M. Fox, and Susan Grandy. "Trends in method of diagnosis of type 2 diabetes mellitus: results from SHIELD." International journal of endocrinology 2009 (2009).

[20] RajeebDey ,VaibhavBajpai, Gagan Gandhi and BarnaliDey, "Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus", IEEE Region 10 and the Third international Conference on Industrial and Information Systems,Dec 2008.

[21] BalakrishnanSarojini, RamarajNarayanaswamy, Nickolas Savarimuthu, and Rita Samikannu. "SVM ranking with backward search for feature selection in type II diabetes databases." In Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, pp. 2628-2633. IEEE, 2008.

[22] Liang Lily, VinayMandal, Yi Lu, and Deepak Kumar. "Multi-dimensional cluster misclassification test for pathway differential analysis of diabetes." In Computer and Computational Sciences, 2007. Second International Multi-Symposiums on, pp. 84-91. IEEE, 2007.

[23] Aekplakorn, Wichai, PongamornBunnag, Mark Woodward, PiyamitrSritara, SayanCheepudomwit, SukitYamwong, Tada Yipintsoi, and RajataRajatanavin. "A risk score for predicting incident diabetes in the Thai population." Diabetes care 29, no. 8 (2006): 1872-1877.

[24] Lindström, Jaana, and JaakkoTuomilehto. "The diabetes risk score." Diabetes care 26, no. 3 (2003): 725-731.

[25] Tuomilehto, Jaakko, JaanaLindström, Johan G. Eriksson, Timo T. Valle, Helena Hämäläinen, PirjoIlanne-Parikka, SirkkaKeinänen-Kiukaanniemi et al. "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance." New England Journal of Medicine344, no. 18 (2001): 1343-1350.

[26] Pan, Xiao-Ren, Guang-Wei Li, Ying-Hua Hu, Ji-Xing Wang, Wen-Ying Yang, Zuo-Xin An, Ze-Xi Hu et al. "Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance: the Da Qing IGT and Diabetes Study." Diabetes care 20, no. 4 (1997): 537-544.

[27] G.V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", journal, Wiley for the Royal Statistical Society, 1980.