International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Classification Framework for Making Decisions on Diabetes Data Trials

[1]Yousif Hamad Efan, [2] Qays Neamah Ibrahim, [3] Ahmed Mutar Awad

Submitted: 14/11/2022 **Accepted**: 18/02/2023

Abstract: One of the most prevalent metabolic conditions that raise blood sugar is diabetes. The complexity of this condition's interdependence on different circumstances makes early detection difficult. To aid medical professionals in the diagnosing process, crucial decision support systems must be developed. This study suggests creating a type 2 diabetes predictive model with high classification accuracy. The study was divided into two main categories. The proposed data mining model can be viewed as to consist of distinct phases namely classification and clustering. Two Classification phase includes the analysis of Data Mining Classification Algorithms and identifying the best algorithm for the health dataset. Classification is performed on Indian Pima Diabetes Dataset. The clustering framework is a method to identify groups within patient records and to detect clusters of attributes that can be used to identify the target class. This is the proposed data mining model that can detect clusters and classes in a given patient record. For achieving better accuracy, clusterization is performed on second dataset i.e., Indian Patient Liver Dataset. The study suggests adopting a bigger population sample without geographic restrictions as a potential solution. Also, it would be interesting to investigate tuning on top of normalised data to further improve accuracy because the created SVM model did not undergo any special hyperparameter tweaking.

Keywords: *Data mining, classification, Patient Liver Dataset, clustering framework*

1. Introduction

Data mining is the technology that allows the discovery of hidden knowledge in accumulated data (Two Crows 1999). Recent developments in data mining have seen a significant shift in research efforts. These efforts were aimed at improving prediction and classification accuracy. Only specific problem areas that have some previous knowledge or can be exploited to enhance classification performance have shown that continuous development of more advanced classification models using commercial software packages is useful (Goebel, Gruenwald, 1999).However, the research that has been done on related data mining techniques shows that they are not able to solve all types of classification problems (Caruana, Mizil). (2006); Han and Kamber 2001. These flaws have led to the development of carefully designed evaluation strategies for data mining classification models.

Data Mining is a relatively new application in healthcare. This research aims to demonstrate that data mining can be applied in healthcare databases to predict and classify data with reasonable accuracy. The learning algorithms need to be given a training set, from which patterns or rules can be extracted to classify the test dataset. This will allow them to make accurate predictions or classifications.

This work will use a variety of data mining algorithms to highlight the benefits and drawbacks. A pre-processing tool is one of the tools. Pre-processing tools are used to convert raw data into an understandable format for the data-mining algorithm. All other tools require data to send to them in different formats.

After testing data has been classified with reasonable accuracy, the classification rules can be extracted and used to make better predictions. This research aims to assess data mining tools for medical and health care applications. It can aid in making timely and accurate decisions.

The proposed data mining model is composed of two distinct phases, namely classification or clustering. Two Classification phases include the analysis of Data Mining Classification Algorithms and the identification of the best algorithm for the health data. The Indian Pima Diabetes Dataset is used for classification.

Clustering frameworks are used to identify groups within patient records and clusters of attributes that help to identify the target class. Figure 1 shows the proposed data mining model for detecting classes and clusters in

yousifxyz@gmail.com
Directorate of education in anbar, Iraq
kaisn2136@gmail.com
Directorate of education in anbar, Iraq
amaacs2@gmail.com
Directorate of education in anbar, Iraq

patient records. Clusterization is used to improve accuracy. This is the Indian Patient Liver Database.

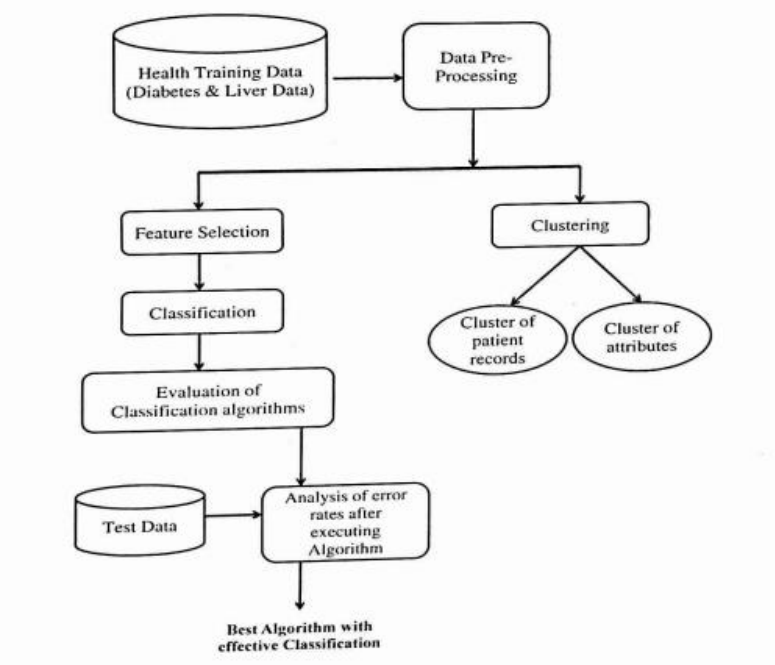

Fig. 1: classification model for clustering health data set

2. Literature Survey

Bala Sundar et al. (2012) Developed a Data Clustering Algorithm to Predict Heart. The results of using K-Means Clustering to predict heart disease were presented in this research paper. K-Means The goal of clustering, a cluster analysis technique, is to divide data into k groups, each of which contains the observation that belongs to the cluster with the closest mean. Each cluster is given a random number of clusters-k, and each cluster starts with random initialization. With this method, the object group was separated into K groups depending on its attributes. This is done by applying the Euclidean distance formula to minimise the square of the distance between the data and the matching cluster centroid. According to research findings, clustering integration gave outcomes that were highly accurate and robust.

Ahmad, A. Mustapha (2011) In this study, multilayer neural networks and tree-based algorithms are compared for accuracy. On the Pima Indian diabetes data set, in particular, the ID3 algorithm and J48 algorithm were applied. The categorization experiment detects whether a patient has diabetes or not with the aid of WEKA algorithms. In comparison to the multilayer perceptrons' accuracy of 81.9%, the results showed that the pruned J48 tree had a greater accuracy of 89.3%. The number of times pregnant attribute was removed, increasing the trimmed J48 tree's prediction accuracy to 89.7%. Zahed Soltani (2016) Diabetes is a major health problem, that can cause physical disability or even death. Methods with the lowest error rate are required to better diagnose this deadly disease. This disease can be diagnosed using different models of artificial neural networks. We have therefore used probabilistic artificial networks to diagnose type II diabetes. We used the 768-sample Pima Indians Diabetes dataset for our investigations. This dataset demonstrates how PNN may be used in MATLAB. By training and testing the Pima Indians' diabetes data, this study also hopes to increase the accuracy of diagnosing type II diabetes. The suggested method's training accuracy and testing accuracy are 89.56% and 81.49%, respectively.

L. AlThunayan, N. AlSahdi (2017) to correctly diagnose diabetes, a prediction approach must be trusted. Those in the medical field may benefit from this. Medical data has been mined extensively for useful information in the field of healthcare. The art of analysing data from various angles and combining it to create usable information is known as data mining. The ability to use data mining for disease early detection is well established. It has a lower cost, higher accuracy, and lower error rates. Classification is one of the most widely used methods in medical data mining. Using the Waikato Environment for Knowledge Analysis, we will examine several data mining techniques (WEKA). The best classification method to predict diabetes will be found after analysis of the findings. We employ a variety of performance parameters, including specificity, accuracy, error rate, and sensitivity, to assess the classifier's correctness.

E. Guldogan, Z. Tunc (2020) In this study, type-2 Diabetes Mellitus (DM) is classified, artificial neural network model estimates are compared, and factors relating to the disease are identified using radial based function and multilayer perceptron approaches. Models of artificial neural networks were employed to predict the risk variables for type 2 diabetes mellitus. The accuracy and specificity of the MLP models were 78.1% and 81.2%, respectively, with an AUC of 0.848, a sensitivity of 71%, and a positive predictive value of 61.6. The accuracy achieved with the RBF model was 76.8%, the specificity was 82.11%, the AUC was 0.8113, the sensitivity was 66%, the positive predictive value was 64.6%, and the negative predictive value was 83%. Moreover, an F-score of 65.3% was attained. The two most significant variables to the MLP model after analysing the effects of variables in the Type 2 DM data set were glucose and body mass index, respectively. RBF was generated from skin thickness and glucose.

T. M. Alam, M. A. Iqbal (2019) Diabetes is a prevalent, chronic condition. Early detection of diabetes can help you receive better care. Many people utilise data mining to make early disease predictions. This study makes predictions of diabetes based on important characteristics. Furthermore mentioned is how the various traits relate to one another. A range of approaches are utilised to assess the importance of attributes as well as for grouping, prediction, and association rules mining for diabetes. The approach of principal component analysis was employed to pick important attributes. The approach of principal component analysis was utilised to identify important qualities. Diabetes was predicted using an artificial neural network (ANN), random forest (RF), and K-means clustering. The ANN method had a high accuracy of 75.7% in predicting diabetes. When making treatment decisions, medical practitioners might find it helpful.

3. Methodology and Materials

Dataset used

Throughout the tests, the Indian Diabetes Database is used. The used dataset is accessible from the UCI repository. It seems to have a decent balance of characteristics caused by diabetes and characteristics that cause diabetes. Following are some remarks from the dataset's introduction.

Pima Indians Diabetes Database

a) Original owners: - National Institute of Diabetes , Digestive and

Kidney Diseases

b) Donor of database: - Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Center, RMI Group Leader Applied Physics Laboratory

The Johns Hopkins University, Johns Hopkins Road, Laurel, MD 20707

If the patient exhibits symptoms of diabetes as defined by the World Health Organization (i.e., if the 2-hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if discovered during routine medical care), this is the diagnostic, binary-valued variable that is being investigated. Near Phoenix, Arizona, in the USA, the population resides.

A true prediction between 0 and 1 is made by their ADAP algorithm. Using a cut-off of 0.448, this was converted into a binary choice. The sensitivity and specificity of their method were 76 using 576 training instances.

Attribute No	Description
1	Number of times pregnant
2	Plasma glucose concentration two hours in a glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin (mu U/ml)
6	Body mass index (weight in kg)/ (height in m)^2/)
7	Diabetes pedigree function
8	Age (years)
9	Class variables (0 or 1)
10	

Missing Attribute Values: None

Class Distribution: class value "1" is interpreted as "tested positive for diabetes" class value "0" is interpreted as "tested negative for diabetes".

When case value is 0 the number of instances is 500 and when it is 1 the number of instances is 268.

Table 2: Pima diabetes dataset

Attribute	Mean	Standard Deviation
1	3.8	3.4
2	120.8	32.0
3	69.2	19.4
4	20.5	16.2
5	79.8	115.2
6	32.1	7.8
7	0.6	0.4
8	33.2	11.9

The data is converted from text into art format as part of the pre-processing phase. This is easily readable by the WEKA software. All variables are coded with numeric data types.

Data mining analyses all the diabetic data using different classification approaches. Software for machine learning is called Weka 3.6.6. The WEKA software's home screen is depicted in Figure 5.1. The WEKA software's home screen is depicted in Figure 5.1. The left-hand portion of the screen contains the rows of characteristics. The distributions of several attributes that can be utilised for data mining are shown in bar graphs on the right.

This category focuses on the evaluation of the best method for feature selection. This is achieved by creating subsets of training datasets using both an algorithm (filter method), and a manual approach to feature choice. Every data set created is run against a variety of attribute selection methods, which are available on WEKA tool. The main techniques on which experiment performed includes:

CfsSubsetEval

ChiSquaredAttributeEval

Classifier Subset Eval

Principal Components

WrapperSubset Eval

Gain Ratio AttributeEval

SVMAttribute Eval

The Cross Validation method of "Attribute Selection Mode" is used. The average of 10 cross-validations and 10 iterations to find the true positives in each dataset (selection method) are used to compute the average. To represent the overall performance of each approach to feature selection, the average of all the values from the different learning methods is also computed. These experiments were performed on WEKA default settings using the "Best Search" search method. The common attributes obtained from the experiment is used for the rest of the experiments.

Parameter Setting

Data mining algorithms that have been discussed thus far require input parameters. These parameters are the settings of an algorithm. Many parameters in the most popular algorithms need to be tuned to make a classifier that is efficient. A literature search revealed some methods that can be used to optimize different classifiers.

J48 is the Decision Tree learning algorithm that is being used in this experiment. It is a Java implementation for Quinlan's C4.5 learning algorithm, introduced in WEKA in 1993 by Quinlan (1993). The pruning setting for J48 was altered too false. Because of the noise in the data, no pruning was done during the learning process of the decision tree classification fire. The classifier's performance was negatively affected if the data were not pruned. J48 is the name of the decision tree in WEKA. It is a modified version of C4.5. The output screen for the classifier will display the classification of the data and the confusion matrix, as shown in Figure 2. Details of the WEKA decision tree is described in the previous chapter. The rules can be extracted from the training data and used to create the decision tree. The decision tree is used to classify the test data.

Two parameters are important for Naive Bayes. These parameters are the kernel estimator. This can improve performance if the normality assumption proves to be incorrect. Supervised discretization can also be used to convert numeric attributes into nominal. All of these were set to false default. The Naive Bayes algorithm, a type of concept learning technique, is used. The Bayes Theorem is used to calculate the probability that all classifications in the database will be classified (Thou et. al., 2004). Figure 5.3 illustrates the output of the WEKA software for classification. To estimate the probability of each class, the normal distribution means, standard deviation, and weighted sum for each attribute are calculated. For prediction, the highest probability class is selected. The output of the classifier for Naive Bayes Algorithm (WEKA) is as follows.

Multilayer Perceptron (MLP), with backpropagation, is the Neural Network algorithm that can be found in the WEKA tool. (Patil & Kumaraswamy 2009). The MBP complexity is determined by the number and number of layers hidden. The momentum applied to weights during updating and learning rate, as well as the number of weights that are updated and how often they are updated, the time it takes to train and the number of epochs required to do so have significant impacts on learning and performance. These parameters are used to train the algorithm. The default settings of WEKA for these parameters are shown in Appendix D. Learning rate is set at 0.3. A momentum of 0.2, and 500 epochs.

JRip was the Java implementation of RIPPER in WEKA. This Association Rule was implemented in this thesis. William W Cohen proposed RIPPER, which stands for Repeated Incremental Pruning. This algorithm is used in the classification system. This algorithm is based upon association rules with reduced error pruning (REP), which is a common and efficient technique in decision tree algorithms. Cross-validation is performed over 10 folds using a default setting for WEKA.

The Support Vector Machine is implemented in WEKA utilising the SMO Support Vector Machine (linear, polynomial, and RBF kernel) in conjunction with the Sequential Minimal Optimization Algorithm and the Sequential Minimal Optimization Algorithm. As a default, SVM with a linear kernel is used; however, the options -E 5 -C 10 and -E 5 -C 10 offer an SVM with a polynomial kernel of degree 5 and lambda of 10. Experiments were evaluated using a 10-fold cross validation process once again.

4. Discussion of Results

These data mining classification models were created using data mining tools. Initial data includes 09 attributes and 768 records. To pre-process the data, an algorithm is used to select the attributes. After attribute selection, missing records are identified and sickened with the mean data from the dataset. Data mining techniques such as Artificial Neural Networks, Naive Bayes and Association Rule are used to classify these 768 records. This section contains the results of the experiments. This section discusses, analyses and evaluates the results in relation to the problem. It attempts to provide a comprehensive assessment of the results, and to also compare the findings to those from peer-reviewed studies. These results are based upon experiments that were performed using the Indian Pima Diabetes dataset, which is available at UCI repository. All experiments were performed using WEKA 3.6.6, which is a data mining machine-learning tool.

To assess sensitivity, specificity, and accuracy, a differentiated confusion matrix is created. This matrix represents the categorization results and is known as a confusion matrix. The confusion matrix is shown in Table 5.4.

Table 3: Classification details

Classification Model	Classified as non-diabetic	Classified as diabetic
Decision tree	407	93
Naïve Bayes	422	78
ANN	416	84
Association Rule	415	85
SVM	449	51

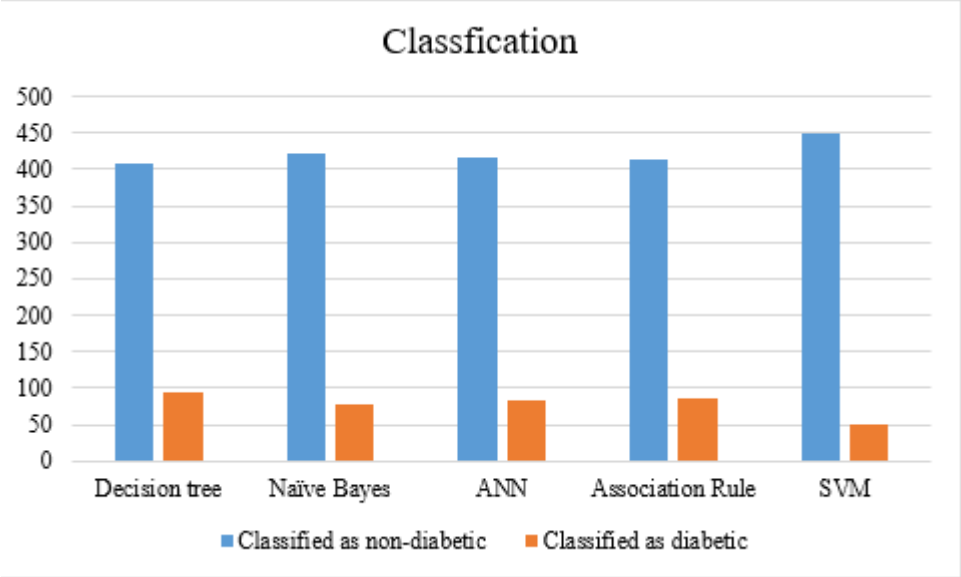

Fig. 3: classification of diabetic and non-diabetic using ML methodologies

Table 3: Performance study of various methods

Model used	True positive Rate	False positive Rate	Precision	Recall	F-Score	ROC Area
Decision Tree (J48)	0.738	0.327	0.735	0.738	0.736	0.751
Naïve Bayes	0.763	0.307	0.759	0.763	0.760	0.891
ANN	0.754	0.314	0.750	0.754	0.751	0.793
Association Rule	0.751	0.317	0.747	0.751	0.749	0.734
SVM	0.773	0.334	0.769	0.773	0.763	0.720

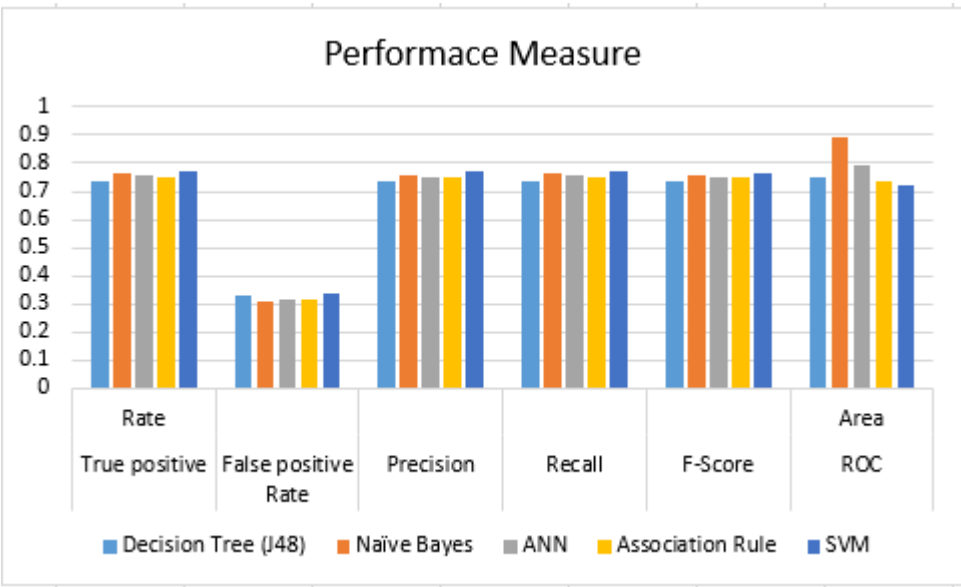

Fig. 4: graphical representation of precision, recall, F-score and ROC-area

Table 3 shows the relative precision and recall for the different techniques. The values of the different techniques varied, as the table shows. Support Vector Machine (SVM), with a 77.3 percent recall, was the best. This means it was able to identify more patients accurately, diabetic and non-diabetic. Decision Tree had a recall rate of only 73.8%. This means that only 73.8% of cases were correctly predicted by the Decision Tree. Naive Bayes was the second-best method to correctly identify patients with diabetes. It achieved a

precision of 74.9 %. The results of the Naive Bayes network were better than those of the Artificial Neural Network (ANN), even though ANN had a ROC (receiver operating characteristic), of 79.3%. The Association Rule Model prediction result was slightly lower than that of ANN. Decision Tree was the most accurate at 73.5 %. Different techniques displayed different characteristics towards the Diabetes Dataset being studied. Figure 4 shows the graphic representation of the models above.

5. Conclusion

This article had the primary purpose of evaluating data mining models to determine which one would offer the best prediction or performance in the context of health informatics. These experiments were performed using the Indian Pima Diabetic Dataset. Initial data on Diabetes are collected and pre-processed. Then, different classification models are used to evaluate the results. It began with the selection of attributes from high-dimensional data sets. Both manual and automatic approaches can be used to select attributes. The results are then evaluated. The best results were obtained using the automatic approach. Data cleaning is a technique that deals with skewed data before the selection of attributes. Next comes the evaluation of models. These metrics are used to evaluate the models: Precision, Recall and Predictor Error analysis. The evaluation process also includes speed measurement.

The five techniques being explored are: Decision Tree, Naive Bayes, Artificial Neural Networks (ANN), Association Rule, Support Vector Machine (SVM), and Decision Tree. Experiments clearly showed that Support Vector Machine (SVM), outperformed all other techniques in terms of Sensitivity Specificity Accuracy and Error Rates. The ROC space point for the SVM Model is also closer to the perfect point (0-1), than other models that show SVM to best predict Diabetes Disease. Naive Bayes was able to predict the outcome faster than SVM, but it is less accurate. Decision Tree performed well at prediction but was slower than SVM. The Neural Network technique is used with Multilayer Perceptron, (MLP). Although its prediction performance was acceptable, it took a long computation time. Association Rule is the fastest technique, but it did not perform well on the prediction.

References

[1] E. Guldogan, Z. Tunc, A. Acet, and C. Colak, "Performance evaluation of different artificial neural network models in the classification of type 2 diabetes mellitus," The Journal of Cognitive Systems, vol. 5, no. 1, pp. 5–9, 2020. View at: Google Scholar

[2] T. M. Alam, M. A. Iqbal, Y. Ali et al., "A model for early prediction of diabetes," Informatics in Medicine Unlocked, vol. 16, 2019.

A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, "Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus," in Digital Information Processing and Communications, pp. 537–545, Springer, Berlin, Germany, 2011.

[3] Z. Soltani and A. Jafarian, "A new artificial neural networks approach for diagnosing diabetes disease type II," International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, pp. 89–94, 2016.

[4] L. AlThunayan, N. AlSahdi, and L. Syed, "Comparative analysis of different classification algorithms for prediction of diabetes disease," in Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing, pp. 1–6, New York, NY, USA, 2017.

[5] International Diabetes Federation, 2020, IDF SEA Members., 2020, https://idf.org/our-network/regions-members/south-east-asia/members/94-india.html.

[6] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," Expert Systems with Applications, vol. 37, no. 12, pp. 8102–8108, 2010.

[7] J. C. Pickup, "Inflammation and activated innate immunity in the pathogenesis of type 2 diabetes," Diabetes Care, vol. 27, no. 3, pp. 813–823, 2004.

[8] D. W. Guthrie and R. A. Guthrie, "Nursing management of diabetes mellitus: a guide to the pattern approach," Home Healthcare Nurse, vol. 23, no. 2, p. 119, 2005.

[9] E. I. Mohamed, R. Linder, G. Perriello, N. Di Daniele, S. J. Pöppl, and A. De Lorenzo, "Predicting Type 2 diabetes using an electronic nose-based artificial neural network analysis," Diabetes, Nutrition & Metabolism, vol. 15, no. 4, p. 215, 2002.

[10] Mayo Clinic Staff, 2020, Prediabetes—symptoms and causes, 2020, https://www.mayoclinic.org/diseases-conditions/prediabetes/symptoms-causes/syc-20355278.

[11] N. C. Sharma, 2019, Government survey found 11.8% prevalence of diabetes in India, 2019, https://www.livemint.com/science/health/government-survey-found-11-8-prevalence-of-diabetes-in-india-11570702665713.html.

[12] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes- estimates for

the year 2000 and projections for 2030," Diabetes Care, vol. 27, no. 5, 2004.

[13] World Health Organization, Global Report on Diabetes, WHO, Geneva, Switzerland, 2016, https://www.who.int/publications/i/item/9789241565257.

A. A. AlJarullah, "Decision tree discovery for the diagnosis of type II diabetes," in Proceedings of the 2011 International Conference on Innovations in Information Technology, pp. 303–307, Abu Dhabi, UAE, April 2011.

[14] U.S. Department of Health and Human Services, Diabetes: A National Plan for Action. The Importance of Early Diabetes Detection, U.S. Department of Health and Human Services, Washington, D.C, USA, 2004, U.S. Department of Health and Human Services, Washington, D.C, USA, 2004, https://aspe.hhs.gov/report/diabetes-national-plan-action/importance-early-diabetes-detection.

[15] M. Komi, J. Li, Y. Zhai, and Z. Xianguo, "Application of data mining methods in diabetes prediction," in Proceedings of the International Conference on Image, Vision and Computing Application, pp. 1006–1010, IEEE, Chengdu, China, June 2017.

[16] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values," IEEE Access, vol. 7, pp. 102232–102238, 2019.

[17] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in Proceedings of the 24th International Conference on Automation and Computing, pp. 1–6, IEEE, Newcastle Upon Tyne, UK, September 2018.

[18] F. G. Woldemichael and S. Menaria, "Prediction of diabetes using data mining techniques," in Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI), pp. 414–418, IEEE, Tirunelveli, India, May 2018.

[19] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[20] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100–107, 2018.

[21] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm-based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in Proceedings of the International Conference on Computing Networking and Informatics (ICCNI), pp. 1–5, IEEE, Lagos, Nigeria, October 2017.

[22] O. Geman, I. Chiuchisan, and R. Toderean, "Application of adaptive neuro-fuzzy inference system for diabetes classification and prediction," in Proceedings of the 2017 E-Health and Bioengineering Conference, EHB, pp. 639–642, IEEE, Sinaia, Romania, June 2017.

[23] S. Ramesh, R. D. Caytiles, and N. C. S. Iyenga, "A deep learning approach to identify diabetes," Advanced Science and Technology Letters, vol. 145, pp. 44–49, 2017.

[24] Marcano-Cedeño, J. Torres, and D. Andina, "A prediction model to diabetes using Artificial Metaplasticity," in New Challenges on Bioinspired Applications, pp. 418–425, Springer, Berlin, Germany, 2011.