Informatics in Medicine Unlocked 16 () 100204

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A model for early prediction of diabetes

Talha Mahboob Alam[a,*], Muhammad Atif Iqbal[a], Yasir Ali[a], Abdul Wahab[b], Safdar Ijaz[b], Talha Imtiaz Baig[b], Ayaz Hussain[c], Muhammad Awais Malik[b], Muhammad Mehdi Raza[b], Salman Ibrar[b], Zunish Abbas[d]

[a] Computer Science and Engineering Department, University of Engineering and Technology Lahore, Pakistan
[b] School of Systems and Technology, University of Management and Technology Lahore, Pakistan
[c] Knowledge Units of Systems & Technology, University of Management and Technology Sialkot, Pakistan
[d] Department of Computer Science, Lahore College for Women University, Pakistan

ARTICLE INFO

Keywords:
Association rule mining
Artificial neural network (ANN)
Data mining
Diabetes
K-means clustering
Random forest

ABSTRACT

Diabetes is a common, chronic disease. Prediction of diabetes at an early stage can lead to improved treatment. Data mining techniques are widely used for prediction of disease at an early stage. In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, and association rule mining for diabetes. Significant attributes selection was done via the principal component analysis method. Our findings indicate a strong association of diabetes with body mass index (BMI) and with glucose level, which was extracted via the Apriori method. Artificial neural network (ANN), random forest (RF) and K-means clustering techniques were implemented for the prediction of diabetes. The ANN technique provided a best accuracy of 75.7%, and may be useful to assist medical professionals with treatment decisions.

1. Introduction

The disease or condition which is continual or whose effects are permanent is a chronic condition. These types of diseases affected quality of life, which is major adverse effect. Diabetes is one of the most acute diseases, and is present worldwide. A major reason of deaths in adults across the globe includes this chronic condition. Chronic conditions are also cost associated. A major portion of budget is spent on chronic diseases by governments and individuals [1,2]. The worldwide statistics for diabetes in the year 2013 revealed around 382 million individuals had this ailment around the world [3]. It was the fifth leading cause of death in women and eight leading cause of death for both sexes in 2012. Higher income countries have a high probability of diabetes [4]. In 2017, approximately 451 million adults were treated with diabetes worldwide. It is projected that in 2045, almost 693 million patients with diabetes will exist around the globe and half of the population will be undiagnosed. In addition, 850 million USD were spent on patients with diabetes in 2017 [5]. Research on biological data is limited but with the passage of time enables computational and statistical models to be used for analysis. A sufficient amount of data is also being gathered by healthcare organizations. New knowledge is gathered when models are developed to learn from the observed data using data mining techniques. Data mining is the process of extracting from data and can be utilized to create a decision making process with efficiency in the medical domain [6]. Several data mining techniques have been utilized for disease prediction as well as for knowledge discovery from biomedical data [7,8].

Diagnosis of diabetes is considered a challenging problem for quantitative research. Some parameters like A1c [9], fructosamine, white blood cell count, fibrinogen and hematological indices [10] were shown to be ineffective due to some limitations. Different research studies used these parameters for the diagnosis of diabetes [11–13]. A few treatments have thought to raise A1C including chronic ingestion of liquor, salicylates and narcotics. Ingestion of vitamin C may elevate A1c when estimated by electrophoresis but levels may appear to diminish when estimated by chromatography [14]. Most studies have suggested that a higher white blood cell count is due to chronic inflammation during hypertension [15]. A family history of diabetes has not been associated with BMI and insulin [16]. However, an increased BMI is not always associated with abdominal obesity [17]. A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision making process. There is a need to combine different parameters to effectively predict diabetes at an early stage. Several existing techniques have not provided effective results when

* Corresponding author.
E-mail address: talhamahboob95@gmail.com (T. Mahboob Alam).

https://doi.org/10.1016/j.imu.2019.100204
Received 26 January 2019; Received in revised form 2 July 2019; Accepted 4 July 2019
Available online 09 July 2019

different parameters were used for prediction of diabetes [18–28]. In our study, diabetes is predicted with the assistance of significant attributes, and the association of the differing attributes. We examined the diagnosis of diabetes using ANN, RF and K-Means Clustering.

2. Related work

Shetty et al. [18] used KNN and the Naïve Bayes technique has been used for the prediction of diabetes. Their technique was implemented as an expert software program, where users provide input in terms of patient records and the finding that either the patient is diabetic or not. Singh et al. [19] applied different algorithms on datasets of different types. They used the KNN, random forest and Naïve Bayesian algorithms. The K-fold cross-validation technique was used for evaluation. Ahmed [20] utilized patient information and plan of treatment dimensions for the classification of diabetes. Three algorithms were applied which were Naïve Bayes, logistic, and J48 algorithms. Antony et al. [21] utilized medical data for diabetes prediction. Naïve Bayes, function-based multilayer perceptron (MLP), and decision tree-based random forests (RF) algorithms were applied after pre-processing of the data. A correlation based feature selection method was employed to remove extra features. A learning model then predicted whether the patient was diabetic or not. Using a pre-processing technique, results were improved when employing Naïve Bayes as compared with other machine learning algorithms. Amina et al. [22] compared different data mining algorithms by using the PID dataset for early prediction of diabetes. Sellappan Palaniappan et al. [23] proposed a heart disease prediction system by using the Naïve Bayes, ANN and decision tree algorithms. Delen et al. [24] used logistic regression, ANN, and decision trees to predict breast cancer using a large dataset. Shadab Adam Pattekari and Asma Parveen [25] developed a web based application for prediction of myocardial infarction using Naive Bayes. Anuja Kumari and R. Chitra [26] used the SVM model to diagnose diabetes using a high-dimensional medical dataset.

3. Methods and materials

3.1. Dataset

The dataset used in this study, is originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases (**publicly available at: UCI ML Repository** [29]). The main Objective of using this dataset was to predict through diagnosis whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification. The Pima Indian Diabetes (PID) dataset having: 9 = 8 + 1 (Class Attribute) attributes, 768 records describing female patients (of which there were 500 negative instances (65.1%) and 268 positive instances (34.9%)). The detailed description of all attributes is given in Table 1.

Our methodology consists of three steps which are explained below.

3.2. Data preprocessing

In real-world data there can be missing values and/or noisy and inconsistent data. If data quality is low then no quality results may be found. It is necessary to preprocess the data to achieve quality results. Cleaning, integration, transformation, reduction, and discretization of data are applied to preprocess the data. It is important to make the data more appropriate for data mining and analysis with respect to time, cost, and quality [30].

3.2.1. Data cleaning

Data cleaning consists of filling the missing values and removing noisy data. Noisy data contains outliers which are removed to resolve inconsistencies [31]. In our dataset, glucose, blood Pressure, skin thickness, insulin, and BMI have some zero (0) values. Thus, all the zero values were replaced with the median value of that attribute.

3.2.2. Data reduction

Data reduction obtains a reduced representation of the dataset that is much smaller in volume yet produces the same (or almost the same) result. Dimensionally reduction has been used to reduce the number of attributes in a dataset [32]. The principal component analysis method was used to extract significant attributes from a complete dataset. Glucose, BMI, diastolic blood pressure and age were significant attributes in the dataset.

3.2.3. Data transformation

Data transformation consists of smoothing, normalization, and aggregation of data [33]. For the smoothing of data, the binning method has used. The attribute of age has been useful to classify in five categories, as shown in Table 2.

Blood glucose concentration in patients who do not have diabetes is different from patients with diabetes. Glucose values have been divided into 5 categories [34] as shown in Table 3.

A strong association has been found between healthy and diabetic patients regarding their blood pressure levels [35]. Blood pressure has been divided into five different categories as shown in Table 4.

The relationship between BMI and diabetes prevalence is consistent. The prevalence of diabetes and obesity is increasing concurrently worldwide. Furthermore, previous studies have shown that BMI is the most important risk factor for type 2 diabetes [36]. BMI values have been categorized into five classes as shown in Table 5.

For the completion of the preprocessing task, selection of significant attributes and transformation of significant attributes into bins are done after data cleaning. The preprocessed dataset visualization is shown in Fig. 1.

3.3. Association rule mining

Data mining techniques are also used to extract useful information to generate rules. Association rule mining is an important branch to determine the patterns and frequent items used in the dataset. It

Table 1
Dataset description and characteristics.

Sr. #	Attribute Name	Attribute Description	Mean ± S.D
1	Pregnancies	Number of times a woman got pregnant	3.8 ± 3.3
2	Glucose (mg/dl)	Glucose concentration in oral glucose tolerance test for 120 min	120.8 ± 31.9
3	Blood Pressure (mmHg)	Diastolic Blood Pressure	69.1 ± 19.3
4	Skin Thickness (mm)	Fold Thickness of Skin	20.5 ± 15.9
5	Insulin (mu U/mL)	Serum Insulin for 2 h	79.7 ± 115.2
6	BMI (kg/m2)	Body Mass Index (weight/(height)^2)	31.9 ± 7.8
7	Diabetes Pedigree Function	Diabetes pedigree Function	0.4 ± 0.3
8	Age	Age (years)	33.2 ± 11.7
9	Outcome	Class variable (class value 1 for positive 0 for Negative for diabetes)	

Table 2
Binning of age.

Age(Years)	Age Bins
≤30	Youngest
31–40	Younger
41–50	Middle aged
51–60	Older
≥61	Oldest

Table 3
Binning of glucose.

Glucose	Glucose Bins
≤60	Very Low
61–80	Low
81–140	Normal
141–180	Early Diabetes
≥181	Diabetes

Table 4
Binning of diastolic blood pressure.

Blood Pressure	Diastolic Blood Pressure Bins
< 61	Very low
61–75	Low
75–90	Normal
91–100	High
> 100	Hypertension

Table 5
Binning of BMI.

BMI	BMI Bins
< 19	Starvation
19–24	Normal
25–30	Overweight
31–40	Obese
> 40	Very Obese

contains two parts: 1) determine the frequent item set, 2) generate rules. An association rule mining approach was developed by Agrawal and Srikan in 1994 which was based on the performance analysis of a Walmart supermarket, buying products with the Apriori algorithm. Association rule mining plays an important role in medical as well as in commercial data analysis to detect and characterize interesting and important patterns. There are several methods to generate rules from data using association rule mining algorithms such as the Apriori algorithm, Tertius and predictive Apriori algorithms. Mostly, association rule-based algorithms are linked with Apriori, which make it a state-of-the-art algorithm. Apriori works as an iterative method to identify the frequent item set in a given dataset, and to generate important rules from it. To determine the association between two item sets X and Y, there is a need to set the minimum support of that fraction of transactions which contains both X and Y called minsupp. The other important task is to set the minimum confidence that measures how often items in Y appear in transactions that contain X, known as minconf, to determine frequent item sets [37]. There were only 268 patients with diabetes in dataset, so only those instances were used to generate rules among them. To develop rules from a given dataset, set minimum support as 0.25 and minimum confidence as 0.9 to generate the following three different rules. Best rules are shown in Table 6.

The association of blood glucose, blood pressure, age, and BMI with diabetes also depended on socio economic, geographic, and clinical factors [38].

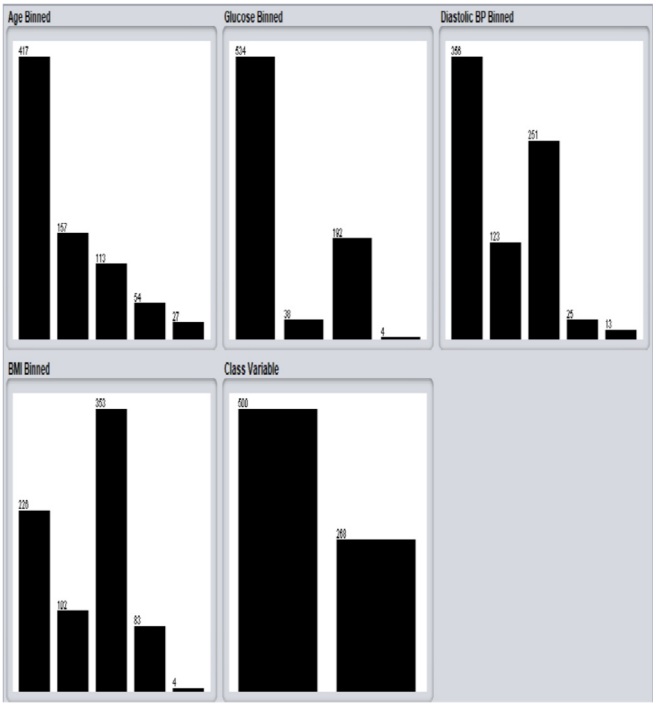


Fig. 1. Pre-processed data visualization.

Table 6
Association Rules using Apriori.

Rule#1. If (BMI = Obesity) → Class = Yes
Rule#2. If (Glucose = Diabetes) → Class = Yes
Rule#3. If (Glucose = Diabetes ∩ BMI = Obesity) → Class = Yes

3.4. Modeling

Three models were used for early prediction of diabetes, following.

3.4.1. Artificial neural network (ANN)

The Artificial neural network (ANN) is a research area of artificial intelligence and an important technique which is used in data mining. The ANN has three layers: input, hidden, and output layer. The hidden layer consists of units that transform the input layer to the output layer. The output of one neuron works as the input for another layer. ANN detects complex patterns and learns on the basis of these patterns. The human brain contains billions of neurons. These cells are connected to other cells by axons and a single neuron is called as perceptron. Input is accepted by dendrites which is taken as stimuli. Similarly, the ANN is composed of multiple nodes that are connected with each other. The connection between units is represented by a weight. The objective of ANN is to convert input into significant output. Input is the combination of a set of input values that are associated with the weight vector, where the weight can be negative or positive. There is a function that sums the weight and maps the result to the output, such as $y = w_1x_1 + w_2x_2$. The influence of a unit depends on the weighting; where the input signal of neurons meets is called the synapse. ANN works for both supervised and unsupervised learning techniques. Supervised learning was used in our study because the output is given to the model. In supervised learning, both input and output are known. After processing, the actual output with compared with required outputs. Errors are then back propagated to the system for adjustment. During training, the data is processed many times, so that the network can adjust the weights and refine them [39].

3.4.2. Random forest (RF)

The random forest method is a flexible, fast, and simple machine learning algorithm which is a combination of tree predictors. Random forest produces satisfactory results most of the time. It is difficult to improve on its performance, and it can also handle different types of data including numerical, binary, and nominal. Random forest builds multiple decision trees and aggregates them to achieve more suitable and accurate results. It has been used for both classification and regression. Classification is a major task of machine learning. It has the same hyper parameters as the decision tree or bagging classifier. The fact behind random forest is the overlapping of random trees, and it can be analyzed easily. Suppose if seven random trees have provided the information related to some variable, among them four trees agree and the remaining three disagree. On the basis of majority voting, the machine learning model is constructed based on probabilities. In random forest, a random subset of attributes gives more accurate results on large datasets, and more random trees can be generated by fixing a random threshold for all attributes, instead of finding the most accurate threshold. This algorithm also solves the overfitting issue [40].

3.4.3. K- means clustering

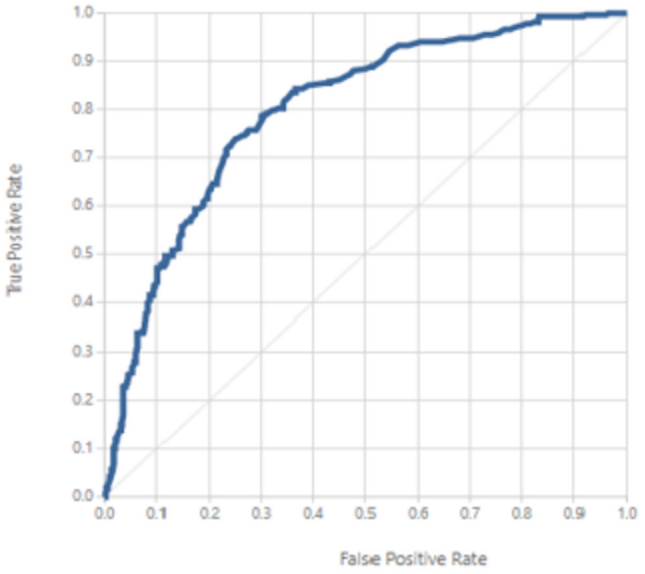
Clustering is the process of grouping similar objects together on the basis of their characteristics. It is an unsupervised learning technique, in which we determine the natural grouping of instances given for unlabeled data. The clusters are similar to each other. However, the objects of one cluster are different from the objects of other clusters. In clustering, intra clustering similarity between objects is high and inter cluster similarity of objects is low. There are many type of clustering, such as partitioning and Hierarchal clustering but in this study, the k-Means clustering method was used. K-Means clustering is relatively simple to implement and understandable, and works on numerical data, in which K is represented as centers of clusters. Taking the distance of each datapoint from the center it assigns each instance to a cluster, and moves cluster centers by taking the means of all the data points in a cluster and repeating until the cluster center stops moving [41].

4. Results and discussion

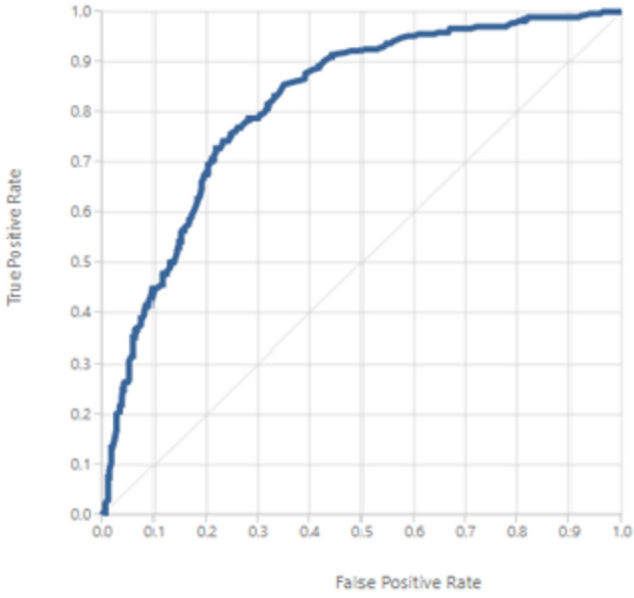
Different classification algorithms were applied on our dataset, and results for all techniques were slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of accuracy and the AUROC curve. The accuracy of models was predicted with the help of a confusion matrix as shown in Fig. 4. First, the random forest algorithm was applied. Experiments were done to tune the model with respect to the number of decision tress and the maximum depth of the decision trees. In the first iteration, the number of decision trees was 8 and the depth of the trees were 4. Again while tuning the model and increasing the number of trees, the results were effective as compared to prior results. Increasing the number of decision trees could be used to obtain improved results, but when the number of trees reached 50, performance diminished. We obtained a best accuracy of 74.7% and an AUROC curve value of 0.806 when the number of decision trees was 32 and the depth of the decision trees was 4. The AUROC curve obtained by using the random forest method is shown in Fig. 2(A). The complete results of random forest is described in Table 7 and the confusion matrix is also shown in Fig. 4(A).

After the random forest algorithm, the ANN was applied to obtain better results. The model was tuned on the basis of number of hidden neurons, number of learning iterations as well as value of initial learning weights. In first iteration, when number of hidden neurons were 50, number of learning iterations were 100 as well as the value of initial learning weights were 0.1, the model has provided satisfactory results. When the values of the tuned parameters were increased, the results worsened. In the 3rd iteration, the values of tuned parameters were decreased; then better results were obtained as compared to the 1st iteration. In the 4th iteration, results were obtained which were most effective when the number of hidden neurons was 5, the number of learning iterations was 10, and the value of initial learning weights was 0.4. The AUROC curve of ANN is shown in Fig. 2(B), which has a value of 0.816 and an accuracy of 75.7%, calculated from confusion matrix as shown in Fig. 4(B).

The complete results of ANN for all iterations is described in



(A) Random forest: Sensitivity= 0.74, Specificity=0.31



(B) ANN: Sensitivity= 0.75, Specificity=0.29

Fig. 2. Performance of ANN and RF on AUROC curve.

Table 7
Performance evaluation of random forest.

Number of decision trees	Maximum depth of decision trees	Accuracy	AUROC
8	4	74.3	0.798
16	4	74.2	0.799
32	**4**	**74.7**	**0.806**
50	4	74.6	0.799

Table 8
Performance evaluation of ANN.

Number of hidden nodes	Number of learning iterations	Initial learning weights	Accuracy	AUROC
50	100	0.1	74.2	0.799
100	1000	0.6	72.8	0.758
20	50	0.4	75.2	0.803
5	**10**	**0.4**	**75.7**	**0.816**

Table 8.

The K-means clustering method was used after the RF and ANN implementation. To apply K-means clustering in our dataset, we normalized the dataset attributes by using the Min-Max normalization technique. Significant attributes were normalized, having the range of 0–1. K-Means clustering was applied by initially setting the value of K = 2, (as in our dataset only two types of patients exist), one for patients with diabetes and the second for patients without diabetes. When the number of clusters was increased, then accuracy decreased. The K-Means clustering predicted 273 to have a value of 1 (positive) and 495 as 0 (Negative). To evaluate the accuracy of K-means clustering, the results were compared to the target class, which shows 203 instances were classified incorrectly, as noted in the confusion matrix of Fig. 4(C). Both clusters were shown in Fig. 3, in which circles in the image show the incorrect instances.

Incorrectly classified instances were 26.43% which show that the accuracy of K-means clustering method was 73.6%.

Accuracy of the proposed models has been compared. The random forest method provided an accuracy of 74.7%, ANN gave 75.7% and K-means clustering method has given 73.6% accuracy. ANN outperforms other methods, as shown in Fig. 5. ANN is a nonlinear model that is straightforward and used for comparing statistical methods. It is a non-parametric model, while the majority of statistical techniques are parametric and require a higher foundation of statistics. The main benefit of utilizing ANN over other statistical techniques is its capacity to capture the non-linear relationship among the concerned variables [42]. The primary weakness of the random forest method is that numerous trees can make the algorithm slow and inadequate for prediction in real time. This algorithm is quick to train, yet very moderate to make predictions once it is trained. A gradually more precise prediction requires more trees, which results in a slower model. Hence, these are the main reasons leading to ineffective results in our study [43].

5. Conclusion and future work

Machine learning and data mining techniques are valuable in disease diagnosis. The capability to predict diabetes early, assumes a vital role for the patient's appropriate treatment procedure. In this paper, a few existing classification methods for medical diagnosis of diabetes patients have been discussed on the basis of accuracy. An classification problem has been detected in the expressions of accuracy. Three machine learning techniques were applied on the Pima Indians diabetes dataset, as well as trained and validated against a test dataset. The results of our model implementations have shown that ANN

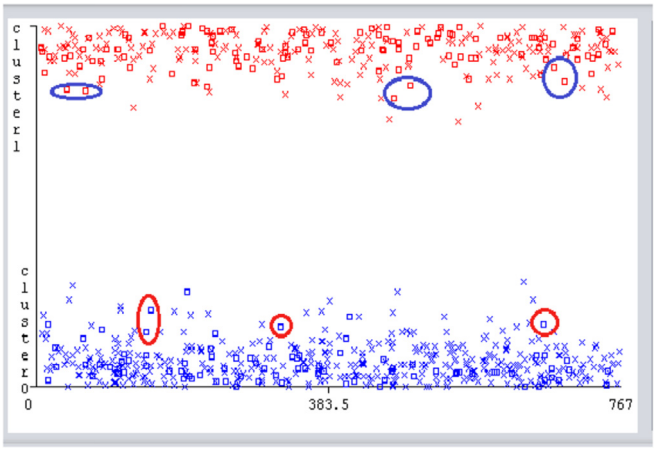


Fig. 3. Correct and incorrect clustered instances.

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive **132**	False Negative **136**
	False	False Positive **59**	True Negative **441**

(A) Random Forest

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive **176**	False Negative **92**
	False	False Positive **95**	True Negative **405**

(B) Artificial Neural Network

		Predicted	
		Positive	**Negative**
Actual	**True**	True Positive **181**	False Negative **87**
	False	False Positive **116**	True Negative **384**

(C) Clustering

Fig. 4. Confusion matrix of proposed models.

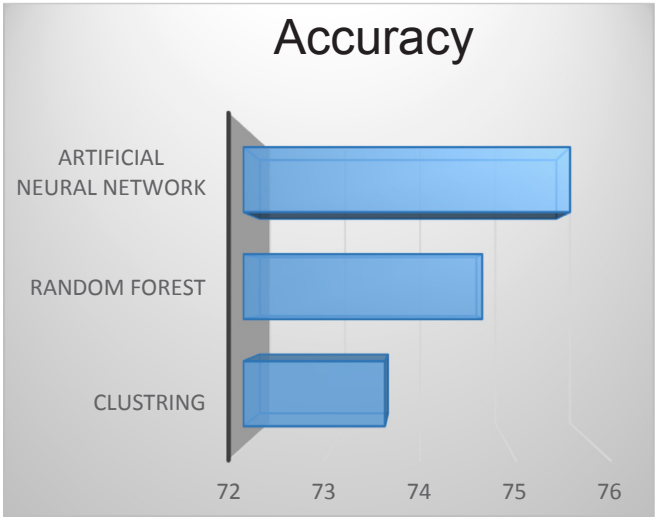


Fig. 5. Accuracy comparison of proposed models.

outperforms the other models. Using association rule mining, the results have shown that there is a strong association of BMI and glucose with diabetes. The limitation of this study is that a structured dataset has been selected but in the future, unstructured data will also be considered, and these methods will be applied to other medical domains for prediction, such as for different types of cancer, psoriasis, and Parkinson's disease. Other attributes including physical inactivity, family history of diabetes, and smoking habit, are also planned to be considered in the future for the diagnosis of diabetes.

Conflicts of interest

There is no conflict of interest.

Acknowledgment

We pay thanks to Dr. Mahdi from AIR University for assistance and guidance especially related to technicalities. We also pay thanks to our respected teacher Dr. Shahbaz who encourage and motivate us.

References

[1] Falvo D, Holland BE. Medical and psychosocial aspects of chronic illness and disability. Jones & Bartlett Learning; 2017.
[2] Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes 2017;66:241–55.
[3] Tao Z, Shi A, Zhao J. Epidemiological perspectives of diabetes. Cell Biochem Biophys 2015;73:181–5.
[4] Organization WH. World health statistics 2016: monitoring health for the SDGs sustainable development goals. World Health Organization; 2016.
[5] Cho N, Shaw J, Karuranga S, Huang Y, da Rocha Fernandes J, Ohlrogge A, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018;138:271–81.
[6] Diwani S, Mishol S, Kayange DS, Machuve D, Sam A. Overview applications of data mining in health care: the case study of Arusha region‖. Int J Comput Eng Res 2013;3:73–7.
[7] Alam TM, Awan MJ. Domain analysis of information ExtractionTechniques. Int J Multidiscip Sci Eng 2018;9:1–9.
[8] Alam TM, Khan MMA, Iqbal MA, Wahab A, Mushtaq M. Cervical cancer prediction through different screening methods using data mining. Int J Adv Comput Sci Appl 2019;10:388–96.
[9] Cobos L. Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat). Endocr Pract 2018;24:43–4.
[10] Dorcely B, Katz K, Jagannathan R, Chiang SS, Oluwadare B, Goldberg IJ, et al. Novel biomarkers for prediabetes, diabetes, and associated complications. Diabetes, Metab Syndrome Obes Targets Ther 2017;10:345.
[11] Singh PP, Prasad S, Das B, Poddar U, Choudhury DR. Classification of diabetic patient data using machine learning techniques. Ambient communications and computer systems. Springer; 2018. p. 427–36.
[12] Negi A, Jaiswal V. A first attempt to develop a diabetes prediction method based on different global datasets. 2016 fourth international conference on parallel, distributed and grid computing. PDGC); 2016. p. 237–41.
[13] Murat N, Dünder E, Cengiz MA, Onger ME. The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels. Biomed Res 2018;29:1370–5.
[14] Radin MS. Pitfalls in hemoglobin A1c measurement: when results may be misleading. J Gen Intern Med 2014;29:388–94.
[15] Merad-boudia HN, Dali-Sahi M, Kachekouche Y, Dennouni-Medjati N. Hematologic disorders during essential hypertension," diabetes & metabolic syndrome. Clinical Research & Reviews; 2019.
[16] Sakurai M, Nakamura K, Miura K, Takamura T, Yoshita K, Sasaki S, et al. Family history of diabetes, lifestyle factors, and the 7-year incident risk of type 2 diabetes mellitus in middle-aged Japanese men and women. J. Diabetes Investig. 2013;4:261–8.
[17] Paley CA, Johnson MI. Abdominal obesity and metabolic syndrome: exercise as medicine? BMC Sports Sci. Med. Rehabil. 2018;10:7.
[18] Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining. Innovations in information, embedded and communication systems (ICIIECS), 2017 international conference on. 2017. p. 1–5.
[19] Singh A, Halgamuge MN, Lakshmiganthan R. Impact of different data types on classifier performance of random forest, naive Bayes, and K-nearest neighbors algorithms. Int J Adv Comput Sci Appl 2017;8:1–10.
[20] Ahmed TM. Using data mining to develop model for classifying diabetic patient control level based on historical medical records. J Theor Appl Inf Technol 2016;87.
[21] Singh DAAG, Leavline EJ, Baig BS. Diabetes prediction using medical data. J Comput Intell Bioinform 2017;10:1–8.
[22] Azrar A, Ali Y, Awais M, Zaheer K. Data mining models comparison for diabetes prediction. Int J Adv Comput Sci Appl 2018;9.
[23] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in Computer systems and applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, 2008, pp. 108-115.
[24] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med 2005;34:113–27.
[25] Pattekari SA, Parveen A. Prediction system for heart disease using Naïve Bayes. Int J Adv Comput Math Sci 2012;3:290–4.
[26] Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. Int J Eng Res Afr 2013;3:1797–801.
[27] Seera M, Lim CP. A hybrid intelligent system for medical data classification. Expert Syst Appl 2014;41:2239–49.
[28] Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Inform. Med. Unlocked 2018;10:100–7.
[29] M. Lichman, "Pima Indians diabetes database," ed. Center for machine learning and intelligent systems.: UCI Machine Learning repository. .
[30] Benhar H, Idri A, Fernández-Alemán J. Data preprocessing for decision making in medical informatics: potential and analysis. World conference on information systems and technologies. 2018. p. 1208–18.
[31] Abidin NZ, Ismail AR, Emran NA. Performance analysis of machine learning algorithms for missing value imputation. Int J Adv Comput Sci Appl 2018;9:442–7.
[32] Liu H, Motoda H. Feature selection for knowledge discovery and data mining vol. 454. Springer Science & Business Media; 2012.
[33] Malley B, Ramazzotti D, Wu J T-y. Data pre-processing. Secondary analysis of electronic health records. Springer; 2016. p. 115–41.
[34] Egi M, Bellomo R, Stachowski E, French CJ, Hart GK, Hegarty C, et al. Blood glucose concentration and outcome of critical illness: the impact of diabetes. Crit Care Med 2008;36:2249–55.
[35] Brunström M, Carlberg B. Effect of antihypertensive treatment at different blood pressure levels in patients with diabetes mellitus: systematic review and meta-analyses. BMJ 2016;352:i717.
[36] Menke A, Rust KF, Fradkin J, Cheng YJ, Cowie CC. Associations between trends in race/ethnicity, aging, and body mass index with diabetes prevalence in the United States: a series of cross-sectional studies. Ann Intern Med 2014;161:328–35.
[37] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Acm sigmod record. 1993. p. 207–16.
[38] Zhenfang X, Zhuansuo W, Qunfang C, Jianjun Y, Xuan ZHANG TY. Prevalence and risk factors of type 2 diabetes in the adults in Haikou city, Hainan island, China. Iran J Public Health 2013;42:222.
[39] Schalkoff RJ. Artificial neural networks vol. 1. New York: McGraw-Hill; 1997.
[40] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2:18–22.
[41] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. IEEE transactions on pattern analysis & machine intelligence. 2002. p. 881–92.
[42] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352–9.
[43] Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip. Rev.: Data Min Knowl Discov 2012;2:493–507.