Review article

Usage of data mining in predicting Diabetes mellitus: A survey

Alaa Aldeen Abd Alwahab **Ahmed**[1,*], Nazim Osman **Bushara**[2]

*1- Faculty of Computer Science and Information Technology, Shendi University, Shendi, Sudan, *alaaldeeen@gmail.com*

2- Faculty of Computer Science and Information Technology, Sudan University for Science and Technology, Khartoum, Sudan

Abstract

The evolution of biotechnology and health medicine has led to a rise in data volume from electronic devices and lab results. Data results are used in data mining to convert this information into knowledge. Data mining can help in make the correct decisions, better diagnosis and excellent treatment, which makes the doctor quickly, defines the problem and best prescribe the treatments. This paper focuses on recent advances in machine learning from 2007 to 2020 that have a significant impact on diabetes detection and diagnosis. Also, it concentrates on the overall survey related to data mining techniques for predicting diabetes and provides a classifying of the current proposals based on the techniques used. In addition, it presents the advantages and disadvantages of those techniques.

Keywords: Data mining, predicting, diabetes, data science.

Introduction

Data Mining is one of the most versatile techniques that have received a warm response from government, healthcare, enterprises and private organizations. It is primarily used for interpreting big data and analytics in order to improve hospital management workflow. Surgeons can benefit from data mining by analyzing large volumes of datasets and gaining relevant insights to perform the operation task more accurately and precisely [1]. It involves some computational process, statistical techniques, classification, clustering and discovering patterns in large data sets. Statistics is the analysis and presentation of numeric facts of data and it is the core of all data mining and machine learning algorithms. It provides analytical techniques and tools to apply on large volume data sets. Statistics include planning, designing, collecting data, analyzing, drawing meaningful interpretation and reporting of the research findings and, due to this, statistics are not only limited to mathematicians, business analysts are also using them. To get the desired output or quantify data statistics uses probability, designing surveys and experiments. Data mining is superior to statistics in that it is more efficient the beginning of science and data science and it covers the entire process of data analysis. While statistics are the foundation and core of the data mining technique, data mining is an exploratory analysis process in which we first explore and acquire data before building a model on the data to find patterns and formulate hypotheses in order to forecast future outcomes or address challenges. Whereas statistics is the confirmative process in which first theories are made and then validation is applied on that theory to test the

datasets, data mining is not very concerned about collection or gathering of data as it is exploratory data analysis. Statistics is more about the collection of data as to get confirmation on the predicted data we need to gather data analyze it to answer questions. The information gathered can be quantitative, qualitative, primary, or secondary. The human body needs insulin to help convert the food it eats into the energy and energy reserves the body needs to function properly. When the body does not make or use insulin properly, a person is considered to have diabetes [2], there are two types of diabetes:

Type1 diabetes used to be called juvenile diabetes. Half of all people with type 1 diabetes are diagnosed during childhood or their early teen years.

Scientists do not know the exact cause of type 1 diabetes. They suspect that it is a combination of factors due to a person's genetics and environment.

Type 2 diabetes tends to develop in people over 40 and used to be called adult-onset diabetes.

Literature review

This section reviews various research works that are related to the proposed work. Some proposed ensemble method to predict Diabetes mellitus at an early stage to save human life, this method found to have a high accuracy of 90.36% [3]. Beside, the suggestion of the Multiple Knot Spline SSVM model for classification problems to evaluate how efficient it was [4]. First, the MKS-SSVM theoretical framework was provided. Then, in order to diagnose diabetic disease, MKS-SSVM was used and compared to SSVM. Findings revealed that the proposed MKS-SSVM performed better, with an accuracy of 93.20%. As well, it has proposed using the UCI repository diabetes dataset and generates decision tree models for classification using LAD tree, NB tree and a Genetic J48 tree [5]; results of the analysis showed that the genetic algorithm achieved the highest accuracy of 98.6%. On the diabetes dataset, it has proposed that a large number of classification algorithms conducted classification with varying degrees of accuracy [6]; the improved weighted least squares support vector machine (WLS-SVM) was used based on quantum particle swarm optimization (QPSO) algorithm to increase prediction accuracy.

Taxonomy based on the technique used

KNN

The K-Nearest Neighbor (KNN) algorithm was used for prediction of diabetes [7]; it has found that the KNN algorithm had the highest accuracy, equal to 70.866%. In similar research focusing on the implementation of K-Nearest Neighbor (KNN) algorithm for prediction of diabetes diagnosis, results showed KNN with a

higher accuracy of 81% [8]. In addition, the K-Nearest Neighbor (KNN) algorithm was employed to analysis and prediction of Diabetes mellitus (DM) patients diabetes and the results computed are found to be very accurate with classification accuracy of 81.77% [9]. As well, the amalgam KNN algorithms were used for classification and prediction of DM [10], a dataset was used from the University of California, amalgam KNN showed the maximum accuracy with 80%, and classification shows better accuracy when the K value is increased to a large value. It has suggested that a diabetes prediction system based on the K-nearest neighbor (KNN) algorithm for removing unnecessary data and reducing processing time [11]. To classify each data sample, a proposed classification approach based on Decision Tree (DT) was proposed. After trials, their proposed approach acquired a high classification result of 98.7%.

j48

A model for an analysis of risk factors for diabetes using a data mining approach has proposed [12], in which Random tree and j48 were used to provide better results to people and doctors regarding diabetes diagnosis, and time taken to build a model for Random tree and J48 was observed to be 0.05 and 0.08 seconds, respectively, but the J48 was selected to be better with a higher accuracy of 80.71%, because in the medical field, accuracy in prediction of diseases is the most important factor rather than the execution time. Also, a study was conducted on analyzing diabetic data using J48 algorithm. This algorithm found to be higher accuracy of 67.16%. The J48 decision tree classifier was used on the data set to build a decision tree model [13]. On the same datasets, an association rule was also applied. The combination of attributes generates a large number of laws. In addition, J48 algorithm was used to build a model on type II diabetes datasets [14]. It has discovered that the J48 algorithm performed better in the supplied dataset, with an accuracy of 93.5%. While some suggested an approach for efficiently predicting diabetes from medical records of patients [15], a modified J48 classifier was used to increase the accuracy rate of the data mining procedure. These experiments revealed that the method outperformed the previous J48 algorithm by a wide margin. The proposed method has been shown to attain an accuracy of up to 99.87%. In addition, a prediction model on the basis of linear regression with the classification method on the real-life diabetes dataset was established [16], the J48 algorithm was used and found to provide higher accuracy of 99.2% [17]. A modified J48 algorithm was implemented to predict diabetes [18], good results for the accuracy of the algorithm has got. This modified algorithm provides 99.87% of accuracy.

SVM

Using machine learning technique Support Vector Machine (SVM) has proposed to predict diabetes [19], it has found that the SVM is the best with a higher accuracy of 90.1%. In a similar research aiming assisting medical institutions, medical research centers and labs with predicting people's DM status [20]. If people are predicted to have a chance of developing diabetes using SVM algorithm, they get SVM with a higher accuracy of 99.50%. In another research, SVMs were proposed for diabetes diagnosis [21]. An additional explanation module was deployed in particular, which converts an SVM's "black box" model into an understandable depiction of the SVM's diagnostic decision. It was observed that SVMs achieved an accuracy of 94% and they found the extracted rules are medically sound and agree with the outcome of relevant medical studies. Also, the SVM was used to classify diabetes in non-laboratory settings [22]. The SVM achieved a higher accuracy of 85% for diabetic patients.

Logistic regression

In a study to analyze the performance of classification algorithms in predicting diabetes, a number of algorithms used, including the Logistic Regression algorithm [23]. It has found that Logistic Regression had a higher accuracy of 80.43%. In similar research, different data mining techniques was used to predict diabetic diseases [24], these prediction models were compared for diabetes or prediabetes incidence using 12 risk factors. It has discovered that the logistic regression model performed best in terms of classification accuracy, with an accuracy of 77.87%.

Random forest

Random forest algorithm was used to build the machine learning model to carry out the diagnosis of diabetes [25]. From the results, Random forest provided a higher accuracy of 100%. In another study, Random Forest algorithm has applied on adult population data to predict diabetic patients using 18 risk factors [26]. This algorithm achieved the best performance, 90% compared to others. Also, the Applied Random Forest algorithm has applied to performance analysis of classifier models to predict diabetes provided better performance and accuracy [27]. Random Forest provided accuracy of 100%.

C4.5

Using data mining techniques such as C4.5, some suggested a model for diagnosing diabetes [28]. C4.5 found to be the best, with a perfect accuracy of 100%. While, another project was to mine diabetes data for efficient classification through mining relationships [29]. The RND TREE Algorithm was found to provide 100% accuracy, but the rule set is large, and this algorithm suffers from data overfitting. C4.5 had a classification rate of 91%. a decision tree was used to build a prediction model for a type 2 diabetes data set [30]. In those models,

it has found that Plasma Insulin attribute is a primary predictor of disease.

C5.0

The performance of Logistic regression, Artificial Neural Networks and C5.0 algorithms were compared for predicting diabetes using common risk factors [31]. It has discovered that the decision tree C5.0 has the best classification accuracy of 80.68%. Also, it has compared the performance analysis of the C5.0 algorithm with other algorithms for predicting diabetes using common risk factors [32]. The decision tree model (C5.0) was given the best classification accuracy.

Naive bayes

Naive Bayes algorithm was implemented to predict diabetes at an early stage [33]. It has found the Naive Bayes classification algorithms are the best at predicting diabetes in primary stage with an accuracy of 76.30%. In a similar study, Naive Bayes and J48 algorithm were employed to find solutions to diagnose diabetes [34]. From results, the Naive Bayes technique found to give the best accuracy of 79.56%.

ANFIS

Different data mining tools were used to measure and investigate the performance of ANFIS, which was found to be the best [35]. Also, Predictive data mining has been used in dose preparation for type 2 diabetes [36]. In addition, the Adaptive Neuro Fuzzy Inference System and Rough Set theory methods were proposed.

ANNs

A system was proposed to improve the detection of diabetic symptoms by synthesizing an intelligent system for retinal image processing by using a multilevel perceptron neural network [37]. From results, it has found that the proposed system provides a value of 98.86%. Also, a study was conducted into detection and prediction of diabetes using machine learning techniques [38]. This used Artificial Neural Networkfor classification of diabetes and discovered that the accuracy of diabetes classification using ANN ranges from 72.2% to 99.51%, whereas diabetes classification using BN ranges from 71% to 99.51%. According to the results, the Bayesian Network attained the maximum accuracy. Some investigators used a large number of machine learning techniques for screening and diagnosis of diabetes using different datasets [39]. Among these algorithms, BP-Artificial Neural Network (ANN) has the best accuracy equaling 99.87%. Neural network techniques with other techniques and Complementary Neural Network classification output were compared [40]; results showed the performance of the Back Propagation Neural Network and Complementary Neural Network techniques are better than other techniques. The best result for dataset used is classified by

Complementary Neural Network, 70.29%, followed by Back Propagation Neural Network at 70%.

SOM

A self-organizing maps (SOM) was employed to depict the responses of patients who had similar responses to survey questions, in order to enable physicians better understand how patients are managing their therapy and where action should be made [41]. Patients who take the correct dose of insulin also likely to administer their injections at the precise time, according to their findings. As well, it has proposed a personalized hybrid model by combining compartmental models and SOM [42]. The model helped patients with type I diabetes to predict their metabolic behavior. Using Self-Organizing Maps (SOM) was proposed to explore the survey data for assessing the self-care behaviors in Type-1 diabetic patients [43]. Specifically, experiments demonstrated that findings through SOM analysis correspond well to the expectations of the clinicians. In addition, these findings inspire the experts to improve their understanding of self-care behaviors for their patients. Another study was conducted to track the features of diabetic women utilizing clustering techniques, as well as approaches clustering and attribute focused induction strategies, and the assessment findings were divided into five groups [44].

Fuzzy logic

Fuzzy Logic is a method of variable processing that allows for the processing of numerous possible truth values using the same variable. Fuzzy logic tries to answer problems using an open, imperfect spectrum of facts and heuristics that allows for a variety of accurate conclusions to be reached. A system that uses a fuzzy logic-based classification technique was innovated to determine the risk factors for prediabetes [45]. The proposed method identified a more accurate set of outlying prediabetes conditions (outliers) among the data associated with a negative diabetes outcome. Also, FCS-ANTMINER has presented, a classification algorithm obtained by combining ant colony optimization and Fuzzy Logic for diabetes disease diagnosis [46]. Results showed that the method achieved the highest classification accuracy among several well-known and recent learning methods, strongly recommending that the combination of Ant Colony Optimization and Fuzzy Logic be used. The application of fuzzy logic to the diagnosis of diabetes was introduced [47], it refers to the ambiguous sets and linguistic variables that play a role in disease diagnosis, especially diabetes. A hybrid classification system has proposed which built from the EM algorithm and fuzzy ID3 algorithm to obtain decision tree for individual clusters [48]. The HCS system was found useful for predicting the disease for binary classification problems.

Other algorithms

Comparatively, a study conducted to systematically collate and review the application of machine learning, data mining methods and supplement tools in the diabetes research field [49], it has found that about 82% of the articles based on diverse supervised machine learning. A personalized diabetes recommendation system called Gluco Guide was developed and tested for Type 2 Diabetes (T2D) patients in order to provide immediate assistance to the patient as well as alerts [50]; a novel data-mining framework called Temporal-Weighted Regression (TWR) was used to discover correlations between recent lifestyle data and BG levels for each patient and the remission rate. Results showed clearly that Gluco Guide is a novel self-management tool for people with Type 2 Diabetes (T2D). Clinical trial showed that Gluco Guide does help T2D patients to alleviate their diabetes conditions based on two standard clinical blood tests. In another research, it has proposed a technique named SAM for prediction of co morbid conditions associated with diabetes [51]. The SAM algorithm, which is an extension of Apriori, is proposed for generating association rules. In a research project to explore significant features of research features to find out the severity of diabetes, CDT, J48, NB tree and REP tree were used to analyze diabetes patient data [52]. Some DT-based models with pruned and unpruned strategies were examined. The classification outcomes wee compared and found out the best classifier based on individual evaluation metrics. CDT (unpruned) showed the best performance than other algorithms, 96.98%.

Results and discussion

By reviewing the papers in recent years, we found that most of the algorithms that predict DM have high accuracy in prediction. Generally, most of the studies used supervised learning, and the rest mainly used relevance analysis. However, the data sets are small in size (less than 10,000 records) used to diagnose diabetes if a person is healthy or not. It was discovered that almost all of the algorithms performed with high accuracy. Although Naive Bayes and SVM have achieved high accuracy, but their shortcomings are limited by the size of the data and they cannot handle large amounts of data, whereas KNN, Logistic regression, j48, and ANN have achieved high accuracy. Also, they have the scalability to handle massive amounts of data. KNN naturally handles multi-class cases and can do well with enough representative data. Neural networks are flexible and good at modeling nonlinear data with a large number of inputs. Logistic regression is easier to interpret, very efficient to train and deals with noise, is unaffected by missing data and it has some

advantages, as shown in Table 2. We also observed that the SOM algorithm was the most commonly used in model building at 16%, as shown in Fig.1.

Table 1. Analysis of different data mining techniques to predict diabetes

Author & year	Dataset	Data mining techniques	Best DM techniques	Accuracy	Performance criteria
Veena Vijayan V. and Aswathy Ravikumar, 2014.	PimaIndian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases.	EM, KNN, K-means, amalgam KNN and ANFIS.	ANFIS	80%	Accuracy
M. Renuka Devi and J. Maria Shyla, 2016.	PimaIndian Diabetic Set from University of California, Irvine(UCI) Repository of Machine Learning databases.	Navie Bayes, MLP, Bayesian Network, C4.5, Amalgam KNN, ANFIS and PLS-LDA, Homegenity-Based, ANN and Modified J48	Modified J48	99.87%	Accuracy
N. Vijayalakshmi and T. Jenifer,2017.	An I.S. nursing home research center in	J48 pruned tree technique and Random tree	J48	81%	Precision
1T.Nithyapriya and S. Dhinakaran, 2017.	Questionnaire	Naïve Bayes, SVM and J48	J48	99.50%	F-Measure
S.Selvakumar1, K.Senthamarai Kannan and S.Gothai Nachiyar, 2017.	PimaIndian Diabetic Set from University of California, Irvine(UCI) Repository of Machine Learning databases.	Binary Logistic Regression, Multilayer Perceptron and KNN	KNN	81%	Sensitivity
Meraj Nabi and Abdul Wahid, 2017.	PimaIndian Diabetic Set from University of California, Irvine(UCI) Repository of Machine Learning databases.	Naive Bayes, Logistic Regression, Decision tree and Random Forest	Logistic Regressio n,	80.43%	Accuracy
[K. Saravananathan and T. Velmurugan, 2016.	A private medical Diabetic center	J48,CART,SVM,KNN	J48	67.16 %	F-Measure
Rashedur M. Rahman and Farhana Afroz, 2013.	PimaIndian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases.	MLP, Naive Bayes and C4.5	Naive Bayes	100%	Accuracy
TahaniDaghistanian dRiyadAlshammari, 2016.	Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia.	SOM, C4.5 and Random Forest	Random Forest	90%	Recall
Minyechil Alehegn , Rahul Joshi and	PimaIndian Diabetic Set from University of	SVM, Naive Net, Decision	PEM	90.36%	Accuracy

Preeti Mulay, 2018.	California, Irvine (UCI) Repository of Machine Learning databases.	Stump, and PEM			
[11] AiswaryaIyer, S. Jeyalatha and RonakSumbaly, 2015.	Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases.	J48 and Naive Bayes	Naive Bayes	79.50%	Accuracy
Xue-HuiMeng , Yi-Xiang Huang , Dong-Ping Rao , Qiu Zhang and Qing Liu, 2013.	Questionnaire	logistic regression, artificial neural networks (ANNs) and decision tree	logistic regression	79.59%	sensitivity
Priya B. Patel, Parth P. Shah and Himanshu D. Patel, 2017.	PimaIndian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases.	Gaussian Naive Bayes, KNN, SVM and Decision tree	KNN	70.86%	Accuracy
K. Rajesh and V. Sangeetha, 2012.	Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases	ID3, C4.5, LDA, Naive Bayes, KNN C4.5	C4.5	91%	Accuracy
Nahla Barakat, Andrew P. Bradley and M. Nabil Barakat, 2010.	National Survey of Diabetes data in the Sultanate of Oman	SVM	SVM	94%	specificity
Tao Zheng, Wei Xie, LilingXu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang and You Chen, 2017.	A regional distributed HER repository ranging	logistic regression, ANNs,C5.0	C5.0	80.68%	sensitivity
Mostafa Fathi Ganjia and Mohammad Saniee Abadeh, 2011.	Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases	FCS-ANTMINER	FCS-ANTMINER	84.24%	Accuracy
Sathees Kumar B and Gayathri P, 2014.	Real time diabetes repository	Simple Cart, J48 and NB Tree	J48	80%	Accuracy
J. Pradeep Kandhasamy and S. Balamurali, 2015.	UCI machine learning data repository	J48, KNN, Random Forest and SVM	Random Forest	100%	Accuracy
Minyechil Alehegnand Rahul Raghvendra Joshi, 2017.	PimaIndian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases.	Random Tree and ANN and SVM	SVM	90.1%	Accuracy

From Table1, we find that most of the previous studies focused on supervised learning using a dataset from the Pima Indian that contains 768 instances and this

volume of data corresponds to some algorithms that are not affected by the size of the data. For example, the Nave Basie and SVM algorithms work well with a small amount of training data. Also, previous studies showed that the j48 algorithm achieved the highest accuracy exceeding 99%, as well as the KNN algorithm achieved the lowest accuracy not exceeding 82%, but they work more with large data. Since the evaluation of the machine learning algorithm for any algorithm is an essential part of the model, the model may give satisfactory results when evaluated using one scale, but give bad results against another, and we note that most researchers measure the performance of the model by accuracy, but some researchers have used recall to measure the accuracy. Sometimes there is a problem of bias (the classifier biases towards a particular class), and this accuracy does not give it, but recall shows whether the classifier has a bias towards a particular class or not.

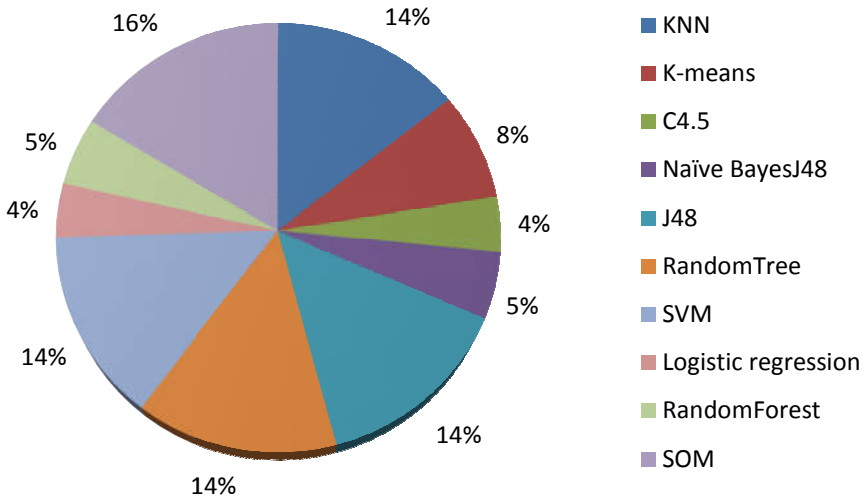


Figure 1. Comparison of varies data mining techniques

It is clear from Fig. 1 that the SOM algorithm was the most used by 16%, followed by KNN, random tree and SVM algorithms. When we talk about the SOM algorithm, we find it easy to explain and understand. Furthermore, lowering the dimensions and using grid clustering makes it easier to see commonalities in the data, but it takes a lot of data to generate meaningful clusters. While KNN is built to handle multi-class cases and can do well in reality with adequate representative data, it is not designed to do so. In addition, when compared to other classifiers, SVM has higher accuracy, and the overfitting problem is less of an issue.

Table 2. Advantage and disadvantage of various techniques

Methods	Advantage	Disadvantage
KNN	It is easy to implement. Training is done in faster manner. Naturally handles multi-class cases. Practically is being good with enough representative data. The best in terms of accuracy compared to other algorithms.	It is requires large storage space. Great search problem to find nearest neighbors.
SVM	Problem over fitting is not like other algorithms. It performs better in high dimensional spaces. When the number of dimensions exceeds the number of models, it is effective. It is memory efficient, as well as utilizes uses a subgroup of training data in the decision function.	It not works well, when we have big dataset because it requires more training time. When the data set has additional noise, it likewise does not perform well. SVM does not directly provide probability estimates; instead, five-fold cross-validation is used in this method.
Logistic regression	Logistic regression is easier to interpret and very efficient to train. It classifies unfamiliar records fairly quickly. It can explicate model parameters as indicators of feature importance.	Logistic Regression should not be used if the number of observations is less than the number of features; otherwise, it may result in overfitting. Using logistic regression to obtain complex associations is difficult.
K-means	Implementation is relatively straightforward. Scalable to big data sets. Assures convergence. It's possible with warm-start the positions with centroids.	Grouping data of various sizes and densities. Clustering outliers. Scaling with number of dimensions.
Artificial neural networks (ANNs)	Neural networks are elastic and can be used for both regression and classification problems. Any data that can be converted to a numerical value can be used in the model. When dealing with nonlinear data with a large number of inputs, neural networks are an excellent choice. When training is done, predictions are very fast. Neural networks can be trained with any number of inputs and layers. Neural networks operate best with a large amount of data.	Neural networks are black boxes, We can't tell how much each independent variable influences the dependent variables, which means we can't tell how much each independent variable influences the dependent variables. Training with standard CPUs is computationally expensive and time demanding. Neural networks based a lot on training data. Over-fitting and generalization become a concern as a result of this. The mode is more reliant on the training data and can be tuned to the data.
Naïve Basie	A Naive Bayes classifier performs better than other models when the assumption of independent predictors is valid. Naive Bayes demands a small size of training data to evaluate the test data. So, the training period is shorter. Naive Bayes is easy simple to use.	Naive Bayes assume all of the qualities to be mutually independent. In actual life, getting a collection of predictors that are totally independent is nearly impossible. If a categorical variable in the test data set has a category that was not included in the training data set, the model will assign a probability of 0 (zero) and will be unable to generate a prediction. This is commonly referred to as Zero Frequency.
J48	An easy algorithm to implement. It deals with noise. It doesn't get much affected by missing data. It can convert the tree to rules.	It does not work well on the small training data set. It is over-fitting. For decision-making, just one attribute is checked at a time.

Random Forest	It's also capable of dealing with continuous features. The Random Forest algorithm, unlike decision trees and other algorithms, is less prone to overfitting. Random Forest algorithm returns the importance of attributes which is a very useful.	Random Forest algorithm may change significantly by a small change in the data. The computations of the Random Forest algorithm can be significantly more complex than those of other algorithms.

Table 2 shows some of the algorithms that were used by researchers to extract the different patterns. To extract patterns, the algorithm that fits with the nature of the problem must be chosen, because each algorithm has a set of pros and a set of cons. For example, The Nave Basie approach outperforms others models, however it necessitates a little quantity of training data in order to estimate the test data. Artificial neural networks are powerful for noise in the training data, can have long training times depending on factors but require processors with parallel processing capability. Plus it black boxes, meaning we cannot know how much each independent variable is influencing the dependent variables.

In conclusion, this paper presented an overview of using the various data mining techniques in predicting diabetes, describing their contributions on this field and providing taxonomy of the technique used. We focused on the capabilities of supervised learning in the prediction of diabetes. In the comparative study among various techniques, the SOM algorithm was the most used by 16% for the prediction. Previous studies showed that the j48 algorithm achieved the highest accuracy exceeding 99%.

References

[1] Kamber I, Frank E. 2011. Data Mining: Practical machine learning tools and techniques, ed: Morgan Kaufmann, San Francisco.

[2] Rude K. 2011. Complete guide to diabetes, 5th ed, American Diabetes Association.

[3] Alehegn M, Joshi R, Mulay P. 2018. Analysis and prediction of diabetes mellitus using machine learning algorithm. *Int J Pure Appl Math* 118: 871–878.

[4] Purnami S, Embong A, Zain J, Rahayu S. 2009. A new smooth support vector machine and its applications in diabetes disease diagnosis. *J Comput Sci* 5: 1003–1008.

[5] Yuvarani S, Selvarani R. 2016. An analysis of decision tree models for diabetes. *Int Res J Eng Technol* 3: 680–684.

[6] Patil B, Joshi R, Toshniwal D. 2010. Hybrid prediction model for type-2 diabetic patients. *Expert sys Appl* 37: 8102–8108.

[7] Patel P, Shah P, Patel H. 2017. Analyze data mining algorithms for prediction of diabetes. *IJEDR* 5: 466–473.

[8] Selvakumar S, Kannan K, Nachiyar S. 2017. Prediction of diabetes diagnosis using classification based data mining techniques. *Int J Stat Syst* 12: 183–188.

[9] Rawat V, Suryakant S. 2019. A classification system for diabetic patients with machine learning techniques. *Int JMath Eng Manag Sci* 4: 729–744.

[10] Vijayan V, Ravikumar A. 2014. Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. *Int J Comput Appl* 95: 12–16.

[11] Kadhm MS, Ghindawi IW, Mhawi DE. 2018. An accurate diabetes prediction system based on K-means clustering and proposed classification approach. *Int J Appl Eng Res* 13: 4038–4041.

[12] Vijayalakshmi N, Jennifer T. 2017. An analysis of risk factors for diabetes using data mining approach. *IJCSMC* 6.

[13] Saravananathan K, Velmurugan T. 2016. Analyzing diabetic data using classification algorithms in data mining. *Indian J Sci Technol* 9: 1–6.

[14] Al Jarullah AA. 2011. Decision tree discovery for the diagnosis of type II diabetes. *International Conference on Innovations in Information Technology*: IEEE 303–307.

[15] Kumar S, Gayathri P. 2014. Analysis of adult-onset diabetes using data mining classification algorithms, *Int J Mod Com Sci* 6: 37–42.

[16] Kaur G, Chhabra A. 2014. Improved J48 classification algorithm for the prediction of diabetes. *Int J Comput Appl* 98: 13–17.

[17] Sohail M. Jiadong R, Uba M, Irshad M, Bilal M, Akbar U, Rizwan T. 2018. Forecast regression analysis for diabetes growth: an inclusive data mining approach. *Int J Adv Res Comput Eng Technol (IJARCET)* 7: 715–721.

[18] Devi MR. 2016. Analysis of various data mining techniques to predict diabetes mellitus. *Int J Appl Eng Res* 11: 727–730.

[19] Joshi R, Alehegn M. 2017. Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. *Int Res J Eng Technol* 4: 426–435.

[20] Nithyapriya T, Dhinakaran S. 2017. Analysis of various data mining classification techniques to predict diabetes mellitus. *Int J Eng Develop Res* 5: 695–703.

[21] Barakat N, Bradley A, Barakat M. 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Tran Inform Technol Biomed* 14: 1114–1120.

[22] Farran B, Channanath A, Behbehani K, Thanaraj T. 2013. Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwaita– cohort study. *BMJ Open* 3.

[23] Nabi M, Wahid A, Kumar P. 2017. Performance analysis of classification algorithms in predicting diabetes. *Int J Adv Res Comput Sci* 8(3).

[24] Meng X-H, Huang Y-X, Rao D-P, Zhang Q, Liu Q. 2013. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaoh J Med Sci* 29: 93–99.

[25] Kumar C, Singh N, Singh J. 2019. Prediction of Diabetes using data miningalgorithm. *Int J Res Appl Sci Eng Technol* 7: 1–8.

[26] Daghistani T, Alshammari R. 2016. Diagnosis of diabetes by applying data mining classification techniques. *Int J Adv Comput Sci Appl* 7: 329–332.

[27] Kandhasamy J, Balamurali S. 2015. Performance analysis of classifier models to predict diabetes mellitus. *Pro Comput Sci* 47: 45–51.

[28] Kumar P, Umatejaswi V. 2017. Diagnosing diabetes using data mining techniques. *Int J Sci Res Pub* 7: 705–709.

[29] Rajesh K, Sangeetha V. 2012. Application of data mining methods and techniques for diabetes diagnosis. *Int J Eng InnovTechnol (IJEIT)* 2: 224–229.

[30] Han J, Rodriguez J, Beheshti M. 2008. Diabetes data analysis and prediction model discovery using rapidminer. Second International Conference on Future Generation Communication and Networking *IEEE* 3: 96–99.

[31] Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. 2017. A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Medical Inform* 97: 120–127.

[32] Steffi J, Balasubramanian D, Kumar M. 2018. Predicting diabetes mellitus using data mining techniques. Int J Eng Dev Res (IJEDR) 6: 460–467.

[33] Sivakumar S, Venkataraman S, Bwatiramba A. 2020. Classification algorithm in predicting the diabetes in early stages. *J Comput Sci* 16: 1417–1422.

[34] Iyer A, Jeyalatha S, Sumbaly R. 2015. Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*.

[35] Rahman R, Afroz F. 2013. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. *J Software Eng Appl* 6: 85–97.

[36] Yıldırım E, Karahoca A, Uçar T. 2011. Dosage planning for diabetes patients using data mining methods. *Pro Comput Sci* 3:1374–1380.

[37] Carnimeo L, Giaquinto A. 2006. An intelligent system for improving detection of diabetic symptoms in retinal images. *IEEE International Conference on Information Technology in Biomedicine, Ioannina* 26–28.

[38] Indoria P, Rathore Y. 2018. ASurvey: detection and prediction of diabetes using machine learning Techniques. *Int J Eng Res Technol (IJERT)* 7: 287–29.

[39] Sun Y-l, Zhang D-l. 2019. Machine learning techniques for screening and diagnosis of diabetes: a survey. *Tehničkivjesnik* 26: 872–880.

[40] Jeatrakul P, Wong K. 2009. Comparing the performance of different neural networks for binary classification problems. Eighth International Symposium on Natural Language Processing: *IEEE* 111–115.

[41] Tirunagari S, Bull S, Kouchaki S, Cooke D, Poh N. 2016. Visualisation of survey responses using self-organising maps: A case study on diabetes self-care factors. *IEEE Symposium Series on Computational Intelligence (SSCI)*: *IEEE* 1–6.

[42] Zarkogianni K, Litsa E, Vazeou A, Nikita K. 2013. Personalized glucose-insulin metabolism model based on self-organizing maps for patients with Type 1 Diabetes Mellitus," in *13th IEEE International Conference on BioInformatics and BioEngineering*: *IEEE* 1–4.

[43] Tirunagari S, Poh N, Hu G, Windridge D. 2015. Identifying similar patients using self-organising maps: a case study on type-1 diabetes self-care survey responses. *arXiv preprint arXiv:1503.06316.*

[44] Padmaja P, Vikkurty S, Siddiqui N, Dasari P, Ambica B, Rao V, Shaik M, Rudraraju V. 2008. Characteristic evaluation of diabetes data using clustering techniques. *IJCSNS,* 8: 244–251.

[45] Rajeswari A, Sidhika M, Kalaivani M, Deisy C. 2018. Prediction of prediabetes using fuzzy logic based association classification. *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*: *IEEE* 782–787.

[46] Ganji M, Abadeh M. 2011. A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. *Exp Sys Appl* 38: 14650–14659.

[47] Radha R, Rajagopalan S. 2007. Fuzzy logic approach for diagnosis of diabetics. *Inform Technol J* 6: 96–102.

[48] Adidela D, Lavanya D, Jaya S, Allam A. 2012. Application of fuzzy ID3 to predict diabetes. *Int J Adv Comput Math Sci* 3: 541–545.

[49] He B, Shu K-I, Zhang H. 2019. Machine learning and data mining in diabetes diagnosis and treatment, in IOP Conference Series: *Mat Sci Eng* 490: 042049.

[50] Luo Y, Ling C, Schuurman J, Petrella R. 2014. Glucoguide: An intelligent type-2 diabetes solution using data mining and mobile computing. IEEE International Conference on Data Mining Workshop: *IEEE* 748–752.

[51] Selvaraj D, Mercy M. 2015. Distributed association rule mining and summarization for diabetes mellitus and its co-morbid risk prediction strategy using FUZZY classifier. *Int J Eng Appl Sci* 2: 58–62.

[52] Howlader K, Satu M, Barua A, Moni M. 2018. Mining significant features of diabetes mellitus applying decision trees: A case study in Bangladesh. *Bio Rxiv* 481994.