Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 1, April 2019, pp.88~95
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i1.pp88-95

Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset

Sinan Adnan Diwan Alalwan
Computer Science and Information Technology College, University of Wasit, Wasit, Iraq

Article Info

Article history:

Received Jul 5, 2018
Revised Oct 3, 2018
Accepted Nov 17, 2018

Keywords:

Accuracy
Classification
Data mining
Diabetes
Self-organizing map

ABSTRACT

Diabetes is a fast spreading illness, which makes to worry millions of people around the globe. The people affected by type-2 diabetes are rapidly increasing and there are no effective diagnostic systems to control the diabetics. As per global health statistics, in western countries, population effected by type 2 diabetics are higher in rate and cost factor for treatment is increasing. There are no effective methods to eradicate the diabetes and it leads to carry out an investigative study on this disease. In existing reviews, researchers are using data analysis approaches to link the cause for diabetes with the patients based on the diet, life style, inheritance details, age factor, medical history, etc. to identify the root cause of the problem. By having multiple key factors and historical datasets, there are some data mining tools were developed, to generate new rules on the root cause of the disease and discover new knowledge from the past data's, but the accuracy was not promising. The main objective of this paper is to carry out a detail literature review and design a conceptual data mining method at initial stage and implement it to improve the result accuracy compared to other classifiers. In this research, two data-mining algorithm were proposed at conceptual level: Self Organizing Map (SOM) and Random Forest Algorithm, which is applied on adult population datasets. The data set used for this research are from UCI machine Learning Repository: Diabetes Dataset. In this paper, data mining algorithms were discussed and implementation results were evaluated. Based on the result performance evaluation, Self-organizing maps have performed better compared to the Random Forest and other data mining algorithms such as naïve Bayes, decision tree, SVM and MLP for diagnosing the diabetes with better accuracy. In future, once system is implemented, it can be integrated with diabetic detector device for faster diagnosis of TYPE 2 diabetes disease.

Corresponding Author:

Sinan Adnan Diwan Alalwan,
Computer Science and Information Technology College,
University of Wasit, Wasit, Iraq.
Email: sdiwan@uowasit.edu.iq

1. INTRODUCTION

Today's fast growing world, there are several health issues arises among the population; one the major illness is diabetes which slowly reduce the lifespan of the individual. The cause of diabetes is due to disorder in metabolism, which leads to higher blood glucose level, heavy urination, continuous thirst, skin dryness and lack insulin segregation. There are three type of diabetes occurs in the body: Type 1 and Type 2 diabetes and another is gestational diabetes which occurs during pregnancy for woman. For this research, type 2 diabetes datasets are taken for analysis and mining. For the concept of data mining algorithms or tool development, the best dataset is diabetes data, which can discover knowledge that is more new and there are specific reasons to carry out the research, they are as follows:

a) Diabetic database will have more past patient data information's
b) Novel knowledge discovery can be made out of the mining which leads less cost for treatment
c) Diabetes will cause side effects if it is not being treated properly, it leads to blindness, failure in kidney, and knee joint pains etc., so it is very important for the doctors to make quick decisions by identifying the significant cases.

Data mining plays an important role in healthcare datasets, where it assists the researchers in healthcare to analyze the datasets, extract the knowledge from complex datasets. With the assistance of new developed tools abd technology, it is very helpful to conduct research in diabetes datasets and improve delivery in healthcare services and improve the accuracy is decision-making process. The data mining methods are widely applicable in various domain, and the methods include pattern recognition, classification approaches, clustering and association methods.

In this research, the comparative result analysis was discussed and shows which data-mining algorithm is better for the classification. To understand the data mining procedures for implementation, need strong fundamentals on the data mining process. The data mining process is divided into five stages, which includes a) Data Cleaning b) Data Preparation c) Data Analytics d) Evaluation e) Visualization.

a) Data Cleaning: Before the process of data mining in diabetic patient datasets, there must be a cleaning process in data. There may be errors, which include missing of values, typo errors, non-related information's and redundancies. In most of the existing works, there are two standard approaches applied to clean the data, which includes format standardization and neighborhood sort approach. The major cleaning in this process will be on reducing the duplications.
b) Data Preparation: In this stage, data preparation cannot take over other stages, data analysis stage should depend on this stage to have a successful result. As per reviews, basic issue is to use whether relational database or flat files. There are two common methods applied, first is based on collection of flat files, which represent all data in the collection set, after that datamining tool is applied for each set of files and outcome is achieved. Another method is data reduction method, where the database is transformed into vector ranks rather considering on individual basis.
c) Data Analytics: In this stage, data mining technology is applied to convert the stored data into effective knowledge. There is various software is available in the market which use multiple models on same data and achieve different results. As per the existing works, Regression tree analyzer is commonly used (i.e. CART software).
d) Evaluation: In this stage, rules extracted from data mining process from the previous stage which may meaningful or not. Therefore, researchers need to undergo a process of evaluation before applying it. Based on this evaluation, results can be improved and hidden patterns can be discovered.
e) Visualization: In this final stage, after extracting the knowledge, need to identify how this knowledge can have used for early prediction of diabetes, therefore in previous stage, tools used may be provide an effective solution or not, but user need to decide how this knowledge on diabetics mining is going to be used in efficiently for prediction.

There are several existing reviews available on data mining research, however data mining in healthcare is very interesting and there were several classifiers are being introduced to improve the accuracy of the mining. Researcher Makinen and his team [1] have applied Self-organizing map in health care data sets to generalize and detect the links or association between the health complications and risk factors for that complications. SOM is used an un-supervised learning framework to cluster the profiles He used 7×12 hexagonal grid mapping with Gaussian neighborhood functions to detect the similarities and variations in the variables. In another research, conducted by author [2] have used the techniques of SOM to study the behavior of type 1 diabetic patients. From his research outcomes, there were suggestions provided to change or adjust the life style to control diabetics. There were various other data mining algorithms applied to different diabetes datasets includes logistic regression, k-nearest neighbor, support vector machines and multifactor dimension reduction, association rules and fuzzy logics [3], [4]. These entire algorithms applied in diabetic datasets have achieved 80% accuracy and specificity. From the past literature reviews, mining algorithms have the potential to predict the diabetes by using public or private datasets, which is smaller or huge size, but still the accuracy is lacking, it will be addressed through enhanced analytics method in this research.

There were also research works carried out on different life style datasets for diagnosing life style disease by applying data mining approaches. In this research, experiments were carried out through datasets from UC Irvine Machine Learning Repository [5]. The most important life style diseases are heart diseases and type 2 diabetes datasets which are reviewed in details in term of algorithms, methods and achieved results. Author Hlaudi Daniel [6] have applied data mining algorithms which includes J48, REPTREE, Naïve Bayes for heart disease datasets, author analyzed each of the algorithms through evaluation conditions which includes Kappa statistics, Mean Absolute Error and Root Mean Square. Accuracy achieved in his

proposed work on J48 is 94.5%, Naïve Bayes is 97.6% and REPTREE is 95.2%. Another research team lead by Shanthakumar [7] have used K-mean clustering algorithm in heart datasets for diagnosing heart attacks, applied algorithm is for pre-processing the data. Mafia algorithm is applied for mining to extract the data. In this research, neural networks are also applied for training the optimal patterns for predicting the heart attacks based on the assigned weights. Accuracy achieved for predicting the heart attack is 78%. Researcher P.K. Anooj [8] have applied fuzzy rule based algorithm on heart disease datasets, by applying the fuzzy approach, knowledge will be retrieved from patient datasets. The proposed algorithm will have two phases, first is a machine based approach to generate fuzzy rules and derive decision tree structure and second phase is creating fuzzy rule for classification and decision support system, the proposed algorithm have achieved a classification accuracy on 90% in hear disease data sets. Researcher Manikandan and her team [9] have used associated rules for extracting the item relations from heart disease datasets, author have e applied MAFIA (maximal Frequent Item set Algorithm) algorithm to achieve higher accuracy of 97%. In this research, datasets were evaluated based on entropy based cross validation and partition approach and each of the results were compared. Datasets includes 19 attributes and results achieved were accurate with high precision values.

Researcher Jayaraman [10] have designed a hybrid classification framework for Prime Indian Diabetic Dataset. The framework consists of two phases, in which first phase includes K-means clustering is applied for identifying and eliminating incorrect classification classes. Second stage applied decision tree C4.5 algorithms by identifying the correct classification class from first phase. The results indicated that rules generated by C4.5 decision tree were cascaded with K-means algorithms, therefore data are categorized for easier interpretation and comparison, accuracy achieved by cascading two phases is 92%. Author Rian Budi Lukmanto and his team [11] have proposed a computational intelligence technique using fuzzy logic for detection of diabetes mellitus, the method proposed by the team is based on acquisition of knowledge and accuracy is achieved is 88%. The research carried by Cheng Hsiung Weng [12] and team on applying different types of neural network for disease prediction, authors have made comparative analysis with single neural network and multiple neural network with diabetic datasets. Secondly, authors have applied statistical testing methods to find out the difference in each of the classifier performance. As per the results, multiple neural network is better than single neural network. Decision tree model was proposed by Kamadi and team [13] to predict the diabetes disease. Based on the proposed work, better decision rules are identified from each of the datasets within the application of fuzzy boundaries. Research achieved by this method is 90%. Research study carried out by Bumi Ju Lee [14] is based on predicting the fasting plasma glucose for diabetic's diagnosis. Author have compared two set of machine learning algorithms which includes logistic regression and naïve Bayes. Results achieved by Naïve Bayes are better than logistic regression.

Researcher Kandasamy [15] have compared various machine learning algorithms to detect the diabetes using data mining approaches. In his research, machine classifiers J48 Decisions Tree, K-Nearest Neighbor, Support Vector are used to classify the patients with diabetes mellitus. Results achieved by Support Vector Machines are better compared to other algorithms. Author Carpenter and his team [16] have applied instant counting algorithm ARTMAP-IC (Adaptive Resonance Theory Match Tracking Algorithm) in diabetic datasets and achieved an accuracy of 80%. The algorithm ARTMAP-IC is modified by ARTMAP search algorithm which allows the network layer to encode the inconsistences in data sets. In this research, predictive accuracy is calculated based on four databases (PIMA, Breas Cancer, Heart Disease and Gallbladder Datasets). Polat and Gunes [17] have applied Principle Component Analysis (PCA) and adaptive neuro fuzzy inference system, based on these techniques, there is an improvement in diabetic's diagnostic accuracy observed by the authors. The proposed system consists of two phases, in first phase, diabetes datasets dimension is identified with 8 features and minimized to 4 features by using Principle Component Analysis (PCA) feature selection process. In second phase, diabetic diagnosis is processed by adaptive neuro fuzzy inference classifier. Datasets used for this study is from UCI machine learning database. The classification accuracy achieved by these methods are 90% compared to other classification approaches.

Author Hasan Temurtas [18] have proposed a neural network architecture trained by Levenberg-Marqaurdt algorithm and probabilistic neural network algorithm for predicting the diabetes disease. Author have compared the results of the study with existing works and have achieved 76% accuracy. Santi Wulan and his team [19] have proposed a Multiple Knot Spline SSVM (MKS-SSVM) approach for predicting the diabetic disease, and achieved an accuracy of 92% for PIMA dataset. Other researchers proposed different classification techniques for diabetes disease includes Genetic programming [20], Generalized Discriminant Analysis approach [21], Feature Selection via Supervised Model Construction (FSSMC) [22]. Research carried out by Srivas [23] is to apply Multiplayer Perception for classifying the diabetic datasets, it is a feed forward network, consists of several layer nodes with single direction connection and trained by back propagation network, accuracy achieved by applying MLP in this work is nearly about 64%. Author Ratna [24] have conducted several experiments on diabetics mellitus datasets by applying

different machine learning techniques and shows the variation in accuracy, but author have not proposed any specific methods to improve the accuracy apart from indicating the strength and weakness of existing works. Researcher Kevin and his team [25] carried out research to identify the depression disorder which overlaps with somatic illness, author have used WEKA tools as part of data mining techniques to diagnose the depression which also related to diabetes, as per the work, results were not promising, still there is research gap to achieve a better accuracy. Based on all the existing works on data mining and machine learning techniques, we have observed that accuracy need to be improved, therefore the proposed work of enhanced Self Organizing Map (SOM) is taken for this research investigation by changing the parameters to improve the accuracy and specificity for diabetes datasets. The problem statement carried out in this paper is to improve the classification accuracy, precision, recall and minimize the feature set by selection and extraction.

The problem statement for this study is to identify whether the given patient is diabetic or non-diabetic patient or pre-diabetic based on the attributes of age, pregnancy count, BMI, blood pressure, insulin, diabetics pedigree and skin fold thickness. In the current research study, all of the attributes are taken as inputs and feed as per SOM data mining algorithm to predict the results. The main objective of the research in this paper is on early detection of diabetes based on chosen attributes and reduce the cost factors on treatments. To early diagnose the disease, patient healthcare data are taken from the repository and raw data is converted into valid information and hidden patterns in the data are discovered as new knowledge, data mining algorithms such as SOM and Random forest carry out the process to design an automated diabetic analytic model.Self-Organizing Map (SOM) is a machine learning method, which is applied in datasets and helpful to analyze the different set of data (i.e. heterogeneous) and delivers a supervised on un-supervised learning models. The concept of SOM is to map the high dimensional datasets, which need to be highly meaningful and helpful to detect the similarities. Thea another classification approach implemented in this research is Random forest, which is compared with SOM for analyzing the accuracy. The rest of the paper is organized as follows: in Section 2, research methodology of data mining algorithm using Self Organizing Map (SOM) and Random Forest Model are explained in detail, followed by implementation and result analysis in Section 3. Finally, the work is concluded in Section 4.

2. RESEARCH METHODOLOGY: DATA MINING THEORETICAL APPROACH

To fulfil the objectives, this study will include two phases as a part of methodology: Selection of Datasets and Attributes b) Data Mining Algorithms: Self-Organizing Maps and Random Forest. *Selection of Datasets and Attributes:* For this work, datasets are collected and validated from Pima Indians Diabetes Dataset (PIDD), which is obtained from the UCI Machine Learning Repository (UC Irvine Machine Learning Repository, 2018). The data sets include patient information in total of 768 female patients, these patients are from the origin of Prima Indian heritage, all of them are woman with age more than 20 years, they were settled near Phoenix, Arizona, USA and considered as resident of USA. In specific, there were 8 important attributes are taken for this research, which are related to individual and medical features and it is indicated with numeric values for each attributes. The output variable will be in binary form where it indicates 0 or 1 with attribute class values. If the class values are interpreted as 0, then it is negative test of diabetes and interpreted as 1, it is positive test of diabetes. In total there are 768 datasets available in the database, while it was analysed for preparation for mining, observed that 500 patients were fall under negative diabetes test and 260 fall under the category of positive diabetes test. Table 1-3 to show the details information on data sets. Original Source of Dataset: National Institute of Diabetes and Digestive and Kidney Diseases and it is stored in UCI Machine Learning Repository. Dataset Investigation: Observing the data includes binary class variables and shows the signs for the patient with diabetes or non –diabetes based the criteria of World Health Organization (WHO), the test is taken based on 2 hours before the meal (i.e. Post load plasma glucose) and glucose level must be at least 200g/ml. Total number of datasets: 768 cases, Total number of attributes: 8 Attribute Missing Values: Yes.

Table 1. Illustrates the Selection of Attributes [5]

Attributes	Description	Values	Range
Age	Age	1	20-80
Pregnancy Count	Number of times of pregnant	2	0-9
Concentration of Glucose	Plasma Glucose Concertation	3	0-199
Blood Pressure	Diastolic Blood Pressure (mmHg)	4	0-122
Serum Insulin- 2 hours	2 hours Serum Insulin (Uml)	5	0-845
Weight- Body Mass Index(BMI)	Body mass in Index in Kg, Height in m-2	6	0-67
Skin Thickness	Skin fold thickness (mm)	7	0-100
Pedigree function of diabetes	Diabetes pedigree function	8	0.08-2

Table 2. Illustrates the Class Distribution

Class Value	Instances
0	500
1	268

Table 3. Illustrates the Statistical Description

Attributes	Values	Mean	SD
Age	1	32.5	10.1
Pregnancy Count	2	3	3
Concentration of Glucose	3	119	34
Blood Pressure	4	67	17.4
Serum Insulin- 2 hours	5	78	112
Weight- Body Mass Index(BMI)	6	31	6
Skin Thickness	7	19.8	16.7
Pedigree function of diabetes	8	0.2	0.3

From above datasets, it is divided into training, testing datasets, therefore 70% of data are used for training, it is constructed to train the model, and 30% are used for testing which is used for testing stage and predict the accuracy. Furthermore, these above attributes are very significant input for data mining algorithms.

Data Mining Algorithms: Self-Organizing Map and Random Forest: Algorithm applied for this dataset is SOM and Random forest algorithm. Self-organizing map is completely full-connected singular layer network, output of this network of arranged in two-dimensional format of nodes. The basic idea of SOM is simple competitive in output layer nodes, it does not have it in single node; all other neighbour nodes are updated. Self-organizing map ability is to learn themselves and detect the patterns, correlation, and predict the output based on the input. In our data sets, SOM act as un-supervised algorithm and supervised algorithm for prediction. R software is used for implementing the SOM algorithm, in which it contains Kohonen package is used, there are two supervised function includes bdf and xyf are used in implementation. SOM has parameters, which are based network size, and training set, these parameters have heavy impact on the performance of the classifier and computing time. The parameters of SOM chosen are GRID, Rlen, Radius. Grid is used for measuring the map size; Rlen is used counting the iteration and Radius is used for measuring the neighbourhood, value may be decreased linearly during training time.

Another algorithm used in Random Forest, it is a supervised learning algorithm, by its name, it will create a forest with more tree and works it out in random manner. There will be a straight relationship between occurrence of trees in the forest and outcome, if there are more trees in the forest, there will be more accuracy, therefore if the overfitting issues arises, it can be solved by random forest algorithm. Apart from that, if there are missing values in datasets, it can also overcome by random forest algorithm. For this diabetic datasets, random forest algorithm is used for implementation through WEKA data mining tool using the set constant parameters and results are achieved.

Evaluation: To categorize the best performance of results through data mining algorithm applied for diabetic datasets, there are two set of standard matrices are being used which are RECALL and PRECISION. Therefore, RECALL is correctly classified diabetic databases and computed by:

$$RECALL = \frac{TRUEPOSITIVE}{TRUEPOSITIVE + FALSENEGATIVE}$$

Precision is based number of relevant classes that are correctly classified and it is computed by:

$$PRECISION = \frac{TRUEPOSITIVE}{TRUEPOSITIVE + FALSEPOSITIVE}$$

True positive is to determine the patients correctly classified as diabetic patients; false negative is to determine, patient who have diabetics but classified as non-diabetic patients; False Positive patient who do not have diabetic and classified as diabetic patients. The learning algorithm will be selected as best based on greater recall and precision.

3. RESULT AND DISCUSSION

Implementation of SOM is performed through R software, Random forest is through WEKA software, and results are achieved. For both the algorithms, there were two metrics carried out, and assessment is processed based on how the model will fit to this unknown datasets and algorithm performance were computed through precision and recall metrics. SOM has achieved precision and recall nearly 92% for the dataset compared with the Random forest with recall and precision of 75%. Table 4 illustrates the results of SOM and random Forest Classifier. The best algorithm chosen is based on higher performance rate based on evaluation metrics. Higher recall and precision values will be considered as the best values of choosing the better algorithm. SOM have achieved higher precision and recall values on the unseen data (i.e. Test data). Comparative analysis of SOM and Random forest algorithm is shown in the Figure 1.

Table 4. Illustrates the Results of SOM and Random Forest Classifier

Data Mining Algorithm	Recall	Precision
SOM- Training Datasets(bdk function)	0.92	0.85
SOM- Testing Datasets(bdk function)	0.90	0.
SOM- Training Datasets(xyf function)	1.0	1.0
SOM –Testing Datasets(xyf function)	0.91	0.83
Random forest Training Datasets	0.75	0.46
Random forest-Testing Datasets	0.65	0.57

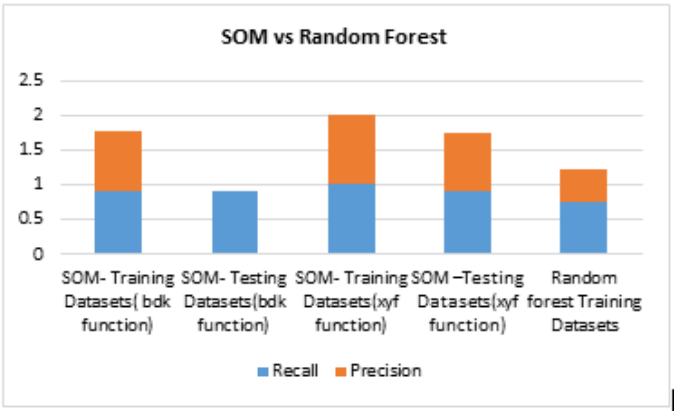

Figure 1. Shows the precision and recall results of SOM and random forest algorithm proposed in the study

The reason for SOM to achieve higher performance compare to random forest is due to constructing the self-models from grid layer, there are multiple layer classification which can improve to achieve greater performance. Therefor this research study SOM and Random forest are applied to forecast the diabetic disease using 8 attributes. Based on the 8 attributes model is constructed and implemented. Therefore, knowledge retrieved this research study based on sample diabetic datasets can be used for larger datasets.

In order to have an effective comparison between other data mining algorithms applied on diabetic datasets through implementation of WEKA and identify which algorithm can work well and suit best in terms of classification accuracy other than random forest, accuracy results are shown in Table 5. There was various research carried out by applying dat mining algorithms over the diabetic's medical datasets, but the research has not shown food accuracy to predict the diabetes, therefore research is more demanding to dig more on domain knowledge and achieve effective medical diabetes diagnosis. Currently in this research, detail comparative study was carried out with proposed SOM work. We have compared the results with random forest and SOM in above Table 4, prediction results were promising, but to make the comparative study effective, more algorithms were applied and results were compared in below tables.

Algorithm used for comparative analysis in Table 5 are a) Gaussian Navies Bayes b) Logistic Regression c) Multiplayer perceptron d) J.48 algorithm e) Support Vector Machine. In Gaussian naïve bayes, each feature is denoted with continuous values and assumption is made to distribute according to the Gaussian distribution. While want to identify the possibilities of people effected towards diabetes, Gaussian naïve Bayes is suitable to identify the positive and negative predictions. In logistic regression, measurement is to find out the relationship between dependent and independent variables based on calculating the

probability using logistic functions. In Multilayer perception, different type of attributes how they are trained and adapted with other attributes to achieve the outcome and reduce the error, so that final result will be the filtered results from each of the neurons. In J.48 algorithm, it is the base work from C4.5 algorithms, in this decision tree algorithm, it will make a decision of which attribute will more decisive and which will the least based on the decomposing the tree in to sub tree. The tree generation will be binary tree and concepts of entropy is used, difference in entropy results will provide the attributes to find out better decisions. In Support Vector Machines (SVM), it is called as discriminative classifier, in which hyperplane separation is carried out and it is designed as linear classification model. Assumptions made in SVM is to plot the training examples in the space, data points will be separated by gap and it will predict the hyperplanes into two classes, focus will be on hyperplane with maximum distance from hyperplane to nearest data pints from any of the classes to find out the possibilities of diabetes. The comparative analysis of graphical representation is shown in Figure 2.

Table 5. Illustrates the Comparative Analysis of Accuracy Over Different Algorithms

Data Mining Algorithm	Accuracy (%)
Gaussian Naïve Bayes	72%
Logistic Regression	75.6%
Multiple Perceptron	76%
J.48 Decision tree algorithm	74%
Support Vector Machine	67%
Self-Organizing Map	85%
Random Forest	78%

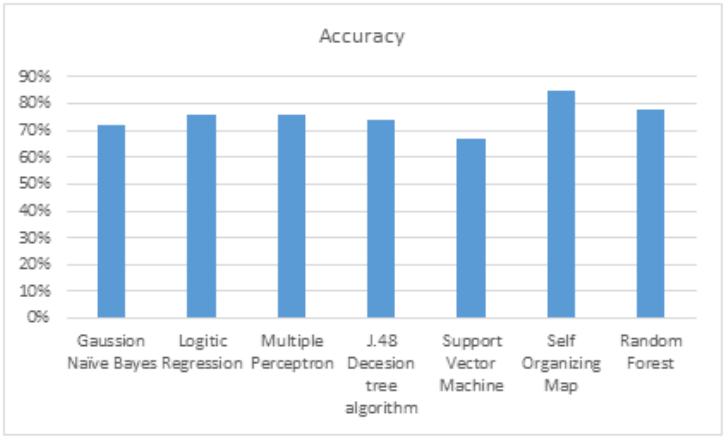

Figure 2. Illustrates the graphical representation of comparative analysis of different algorithms

While comparing the accuracy of the different algorithms, it shows that Self organizing map (SOM) is effective hence its precision, recall and accuracy are better compared to random forest and other data mining algorithms. In Gaussian naive Bayes, results were not promising, but for certain datasets it shows predominant results after set of validation. Decision tree (J.48) has been represented in a graphical images and each of precedence of attributes have calculated the priority attributes and predicts the results within arrange of 74%, compared to SOM it is less accurate but compare to SVM, it shows the better classification accuracy. When compares to logistic, J.48 and naïve Bayes classification, MLP have shown promising accuracy result of 76%. Therefore, compared to other algorithms, SOM have better accuracy of achieving 85% followed by random forest with accuracy of 78% for diabetic datasets. Comparative analysis of applying different data mining algorithm for same diabetes dataset which is able to analyse and predict the accuracy.

4. CONCLUSION

This research aims to propose a two effective data mining algorithms for classifying the diabetic detection; therefore, model designed from data mining algorithms will help the doctors to make faster

decisions in diabetics and applied in any healthcare domain. In this research study, PIMA dataset from UCO machine learning repository is obtained with eight specific attributes for mining. There were two algorithms are proposed, implemented and evaluated which includes SOM (functions using xyf and bdk) and random forest classifier algorithm. The results achieved based on this proposed data mining models can support healthcare service providers for making effective decisions. In future work, this model can be enhanced for controlling diabetes through designing a diabetic control plan, because sometimes difficult to identify diabetes at early stages which leads to last stage and creates problems.

REFERENCES

[1] Ville-Petteri Mäkinen et.al., "Case Report on Metabolic Phenotypes, Vascular Complications, and Premature deaths in a population of 4197 patients with type 1 diabetes", *Diabetes Journal*. Vol 57, No 9, pp.2480-2487, 2008.

[2] Santosh Tirunagari et.al., "Identifying Similar Patients Using Self-Organizing Maps: A Case Study on Type-1 Diabetes Self-care Survey Responses", *arXiv.org e-Print archive.* Vol 0631, No 1, pp.1503-0631, 2015.

[3] Nahla H. Barakat., "Intelligible support vector machines for diagnosis of diabetes mellitus", *IEEE Transactions on Information Technology in Biomedicine*, vol 14, no 4, pp.1114-1120, 2010.

[4] Mostafa Fathi Ganji., "A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis", *Journal of Expert Systems with Applications*, vol 38 no 12, pp. 14650 – 14659, 2011.

[5] UC Irvine Machine Learning Repositor, Retrieved on 10th February 2018 from http://archive.ics.uci.edu/ml/., 2018.

[6] Hlaudi Daniel Masethe, Mosima Anna Masethe., "Prediction of heart disease using classification algorithms", *Proceedings of the world congress on engineering and computer science*, San Francisco, USA, pp. 185-191, 2014.

[7] Shantha Kumar B. Patil., "Extraction of significant patterns from heart disease warehouse for heart attack predictions", *International Journal of Computer Science and Network Security*, vol 9 no 2, pp.101-110, 2009.

[8] P.K. Anooj, "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules", *Journal of Computer Science*. vol 24 no 1, pp. 27-40, 2012.

[9] V.Manikandan, S. Latha.," Predicting the analysis of heart disease symptoms using medical data mining methods", *International Journal of Advanced Computer Theory and Engineering*. vol 2, no 1, pp. 46-51, 2013.

[10] Jayaraman Kkaregowda, A. G, Manjunath.," Rule based classification for diabetic patients using cascaded k-means and decision tree C4.5", *International Journal of Computer Applications*. Vol 45 no.12, pp.45-50, 2012.

[11] Rian Budi Lukamanto, Irwansyah., "The early detection of diabetes mellitus(DM) using fuzzy hierarchical model", *Journal of Elsevier*. Vol 59, pp.312-319, 2015.

[12] Cheng- Hsiung Weng, "Disease prediction with different type of neural network classifiers", *Journal of Elsevier*. Vol 33, pp.277-292, 2014.

[13] Kamadi V.S.R.P, Varma A.," A computational intelligence approach for a better diagnosis for diabetic patients", Journal of Elsevier. vol 40, pp.1758-1765, 2014.

[14] Bum Ju Lee, Boncho N, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing of diabetes", *IEEE Journal of Biomedical and Health Informatics*. Vol 18, no 2, pp.555-561, 2014.

[15] Kandasamy P, Balamurali., "Performance analysis of classifier models to predict diabetes mellitus", *Journal of Elesevier*, vol 47, pp.45-51, 2014.

[16] Carpenter G.A, Markuzon., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", *Journal of Neural Networks*. Vol 11, pp. 323-336, 1998.

[17] Kemal Polat, S. Gunes., "An expert system approach based on principle component analysis and adaptive neuro fuzzy inference system to diagnose of diabetes disease", *Journal of Elsevier: Digital Signal Processing*. Vol 17, pp.702-710, 2007.

[18] Hasan T, Nejat Y, Feyzullah., "A comparative study on diabetes disease diagnosis using neural networks", *Journal of Expert Systems with Applications*. Vol 36, pp. 8610-8615, 2009.

[19] Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain, Rahayu S.P., "A new smooth support vector machine and its application in diabetes disease diagnosis", *Journal of Computer Science*, vol 5,no.12. pp.1006-1011, 2009.

[20] Muhamad Waqar Aslam, Nandi A.K., "Detection of diabetes using genetic programming", *European Signal Processing Conference*, Aalborg, Denmark, pp. 202-212, 2010.

[21] K. Polat, Gunes S, Aslan A., "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine", *Journal of Expert System with Application*. Vol 34, no.1, pp. 214-221, 2008.

[22] Yue Huang, Paul Mccullagh, Norman Black, Roy Harper., "Feature selection and classification model construction on type 2 diabetic patients", *Journal Artificial Intelligence in Medicine*, vol 41, no.3, pp. 251-262, 2007.

[23] K. Srinivas, et al., "Hybrid Approach for Prediction of Cardiovascular Disease Using Class Association Rules and MLP", International Journal of Electrical and Computer Engineering, vol. 6, no. 4, pp. 1800-1810, 2016.

[24] Ratna Patil, et al., "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes", International Journal of Electrical and Computer Engineering, vol. 8, no. 5, pp. 3966-3975, 2018.

[25] Kevin Daimi, et al., "Using Data Mining to Predict Possible Future Depression Cases", International Journal of Public Health Science, vol. 3, no. 4, pp. 231-240, 2014.