IJPSR (2018), Volume 9, Issue 7 (Review Article)

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 11 March, 2018; received in revised form, 24 May, 2018; accepted, 31 May, 2018; published 01 July, 2018

ANALYSIS OF DATA MINING AND SOFT COMPUTING TECHNIQUES IN PROSPECTING DIABETES DISORDER IN HUMAN BEINGS: A REVIEW

Prableen Kaur and Manik Sharma *

Department of Computer Science and Applications, DAV University, Jalandhar - 144012, Punjab, India.

Keywords:

Diabetes, Diagnosis,
Data Mining Techniques,
Soft Computing, Hybrid techniques

Correspondence to Author:
Manik Sharma

Assistant Professor,
Department of Computer Science and Applications, DAV University, Jalandhar - 144012, Punjab, India.

E-mail: manik_sharma25@yahoo.com

ABSTRACT: Diabetes is one of the deadliest and non-contagious diseases that can adversely affect several parts of human body. Early prognosis of diabetes can inkling the grievous complications and help to save human life. Several researchers have used different data mining (Iterative Dichotomiser 3, Random Forest, Support Vector Machine, k-Nearest Neighbour, C4.5) and soft computing (Genetic Algorithm, Ant Colony Optimization, Particle Swarm optimization, Artificial Bee Colony) techniques to prospect diabetes in human beings. In last 10 years, C4.5 was the most preferred choice for mining diabetic patients. Likewise, in soft computing, maximum number of researchers have used genetic algorithm. Furthermore, the usage of pre-processing techniques is significantly increasing in diabetes diagnosis. It is also observed that rate of accuracy achieved in diagnosing diabetes using traditional data mining lies in 68.5% - 95.3%. Likewise, the range for soft computing and their hybridized use lies in 74% - 100%. In addition, rate of accuracy achieved using GA based hybridized approach is better than the accuracy obtained using PSO as well as ABC. Most of the researchers have used textual and numeric data for diabetes diagnosis. Few researchers have used images for the same. However, no significant research is found where diabetes has been diagnosed using audio or sound. Moreover, the diagnostic results obtained using image based data are not as good as obtained using textual or discrete data. Therefore, an attention is still obligatory to develop smart diabetes diagnostic system that can effectively work on different types of data like text, images as well as sound.

INTRODUCTION: Diabetes is a persistent metabolic and pancreas affected human disorder. The devastating rise in obesity and sedentary life style has made it as a universal epidemic [1, 2]. Numbers of individuals are getting affected by different types of diabetes. Human body changes most of the food into glucose and it is the primary source of energy in the body. Pancreas assists in digestion and releases a hormone called insulin that transforms glucose into body cells.

QUICK RESPONSE CODE

DOI: 10.13040/IJPSR.0975-8232.9(7).2700-19

Article can be accessed online on: www.ijpsr.com

DOI link: http://dx.doi.org/10.13040/IJPSR.0975-8232.9(7).2700-19

The failure of pancreas to produce sufficient amount of insulin or production of insufficient insulin may lead to excessive amount of glucose in the body. Due to excessive glucose in blood, BSL (Blood Sugar Level) increases that ultimately leads to diabetes [3].

It is one of the deadliest diseases that claim millions of lives each year. According to the WHO (World Health Organization), it was estimated that 3.4 million deaths are caused due to high blood sugar. It has been found that the over diagnosis of diabetes may lead to comorbidity like cognitive impairment, stroke, cancer, kidney problem *etc*. Therefore, it should be diagnosed at the earliest. In year 2000, India topped the world with 31.7 million people suffered from diabetes followed by China with second place and United States with third

place [4]. It is predicted that by the year 2030 diabetes mellitus may affect up to 79.40 million people in India [5]. In last 40 years, a fourfold rise has been witnessed for this contagious disease [6]. According to International Diabetes Federation, in 2017, there are around 425 million populations suffering from diabetes across the world. It is also estimated that by 2045 the raise in the diabetic population will be increased by 32% [7]. Currently, China, India, USA, Brazil, and Russia are the top five countries with highest rate of diabetic population [8].

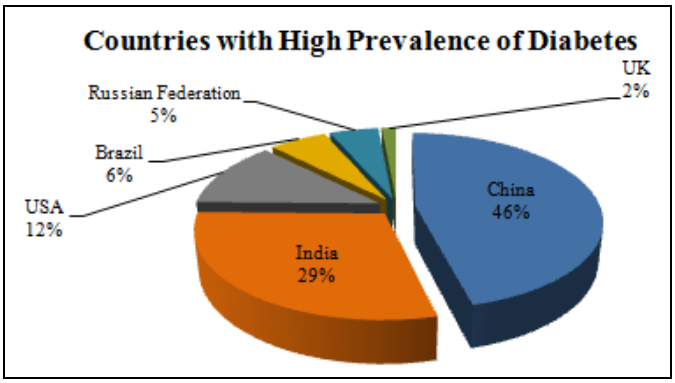

FIG. 1: NUMBER OF PEOPLE AFFECTED WITH DIABETES

There are five different types of diabetes named as Type 1, Type 2, Gestational, Type 4 and Type 5. Type 1 diabetes is Insulin Dependent Diabetes Mellitus (IDDM) caused due to the damaged cells that maintain the optimal level of insulin. Therefore, pancreas becomes unable to produce required level of insulin for human body. To control the sugar level, insulin must be injected every day by taking shots of insulin that helps to control the BSL. It is less common form of diabetes usually diagnosed in children and young adults, hence also called juvenile diabetes. Overall 10 - 15% people affected with type1 diabetes [3, 9, 10]. Type 2 Diabetes is adult-onset diabetes or Non-Insulin Dependent Diabetes Mellitus (NIDDM). It is a common disorder that develops in the middle and old aged people. About 90 - 95% of diabetic patients around the world suffer from type 2 diabetes. Patients with this type of diabetes are renitent to the action of insulin. To control BSL, type2 diabetes patients are recommended to follow routine exercise and have healthy diet. Sometimes small amount of medication is also suggested to some patients [9, 11]. Pregnant women are general candidate of gestational diabetes.

The disease endangered highly in the women above 30 years age or the women gaining extra weight during pregnancy. However, only 5 to 11% pregnant women are affected by this adhoc type of disease. The Gestational diabetic patient's average reading for FPG is greater than 90 mg/dl and 1 h postprandial glucose is greater than 120 mg/dl. When pregnancy is over, the exigency of insulin in body returns to normal and the diabetes resolves. Nevertheless, the women affected with gestational diabetes are more prone to diabetes in future [3, 9]. The Type 3 diabetes is related to Alzheimer's patients and is arises due to the inability of neuron in brain to react to insulin in brain. The patient suffering from this type of diabetes requires insulin in body within two months of diagnosis. According to World Alzheimer Report, 46.8 million people are suffering from Alzheimer globally [12]. Type 4 diabetes has been recently introduced by Salk Institute researchers in November, 2017. Patients suffering from type 4 diabetes had high level of immune cells inside the fat tissues. Whereas, one with type 2 diabetes had low level of immune cells inside fat tissue, but has more fat tissue in body. It normally occurs in old age persons as it is the age associated insulin resistance [13]. **Fig. 2** represents the risk factors and symptoms associated with five types of diabetes.

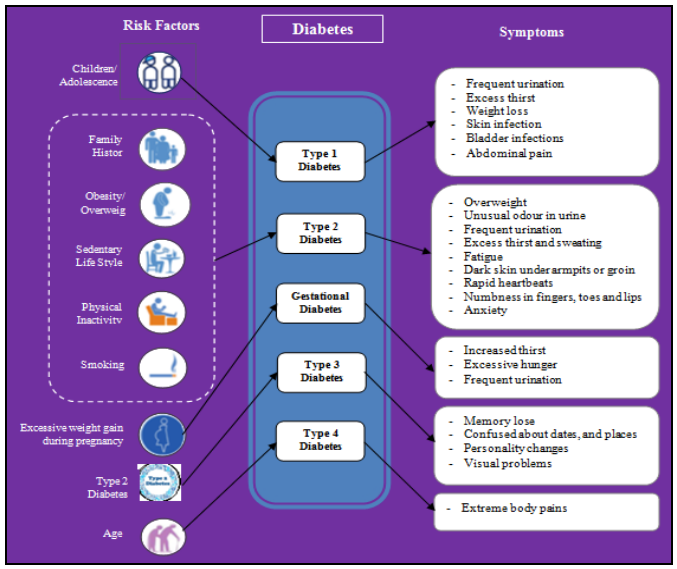

FIG. 2: RISK FACTORS AND SYMPTOMS ASSOCIATED WITH DIABETES

The data scientist have employed different data mining and soft computing techniques to mine and diagnose the diabetic patients data which are cheaper in terms of time as well as costs.

This study is carried out to analyse the role and performance of different data mining, soft computing and hybrid approaches in diagnosing diabetic patient's datasets. The study is prophesied to answer the following queries:

1. What are the distinct types of diabetes mellitus that impinge on the human body?
2. What are the symptoms and risk factors related to diabetic patients?
3. Does the prevalence of diabetes escalate with the increase in age of individual?
4. What type of data mining techniques helps to determine this disease with the best viable rate of accuracy?
5. What are the common features in the dataset that help to diagnose diabetes in patients?
6. What is role of soft computing techniques in diagnosing the different stages of diabetes?
7. What is the effect of data pre-processing on data classification?
8. Does the hybrid algorithm predict diabetes with better accuracy than simple data mining algorithm?

Fig. 3 shows how diabetes affects human life. When this fatal disease is not managed or cure on time will leads to the acute complications over human health, wealth and life. Uncontrolled diabetes can damage blood vessels that further increase the risk of cardiac disorders, blindness, kidney failures, and even cause the premature death. Expenditure for treating and preventing diabetes imposes the economic burden on human. According to WHO report in 2016, it was estimated that the total expenditure on treating diabetes in world was US$ 827 billion annually [9].

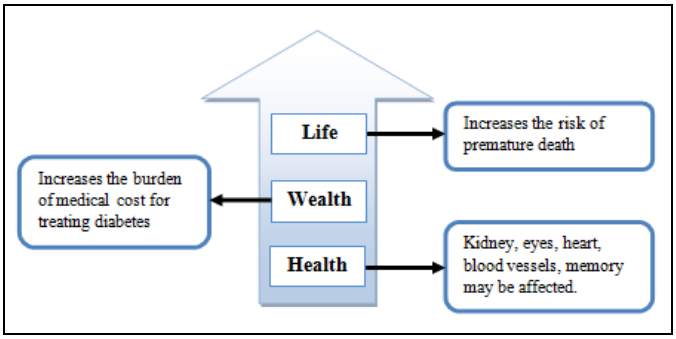

FIG. 3: EFFECT OF DIABETES ON HUMAN BEINGS

The patients' hospital readmission rate significantly increases the healthcare cost. There are two types of patients' readmission *i.e.* readmission within or after 30 days [14]. To reduce the readmission rate of diabetic patients one should take care of parameters *viz.* selection of discharge regimen, acknowledgement of diabetes diagnosis, glycemic management teams, and post discharge support [15]. M.D Silverstien *et al*., have been predicted in their study that the patient with age greater than 65 year and males have under more risk for 30 days readmission [16].

Related Work: Miroslav Marinov *et al*., (2011) reviewed 31 articles related to diabetes diagnosis. The study was sub grouped in classification, clustering and association data mining methods. Authors stated that data mining will have a bright future in biomedicine. However, the precise classification accuracy comparison was missing [17]. Preeti Verma *et al*., (2016) reviewed various studies with classification techniques for diabetes diagnosis. The results showed that Support Vector machine (SVM) effectively classify diabetes disorder. The rate of accuracy achieved using SVM is 96.58%. Authors have not explored the effect of data pre-processing effects on the predictive accuracy of diabetic patients [18].

Ioannis Kavakiotis *et al*., have performed a systematic review of 103 articles for the prediction and diagnosis of diabetes using data mining and machine learning. The articles were categorized into five sections *viz*. biomarker identification, diabetes complications, therapies and drugs consumption, genetic factors and disease management. Authors have shown that the results produced by SVM are better than other techniques [19].

Anjali Khandegar presented a review to construe different data mining techniques to predict diabetes. The study has shown standards to analyze the patients' behaviour and life style parameters such as emotions, physical activities, eating habits *etc*. The retrieved information can be employed to examine clinical parameters, other disease expectancy and treatment planning. However, accuracy comparison of different methodologies has not been mentioned [10]. Tushar Deshmukh *et al*., reassess the study of different authors to signify the role of data mining techniques for diabetes diagnosis. Authors have highlighted the efforts of different authors in analysing the different types of diabetes.

The study concluded that, by predicting diabetes at early stage save patients to suffer from other health issues [20]. F. M. Okikiola *et al*., have shown a systematic review of different data mining techniques used for diabetes diagnosis. Authors have selected 37 articles in their study from fifteen different sources. The main emphasize on this survey was on the diabetes diagnostic management system. In this paper, authors have extensively mentioned the inclusion and exclusion criterion. Authors mentioned the main contribution of the different authors along with physical, nutritional and drug prescription based recommendation. However, the basic details about the diabetes and the mining techniques employed by different authors have not been explained. In addition, authors have not mentioned the predictive rate of accuracies in mining diabetes [21]. S. Pawar *et al*., presented a very brief review on diabetes diagnosis. In the title authors mentioned "An extensive survey". However, it is a very brief and covering only 28 research article. In this paper, authors simply summarize the contribution of different researchers. However, again the detail about the dataset, pre-processing techniques, inclusion and exclusion criterion was missing [22].

Key Points of this Study are:

- ✓ Diabetes and its types have been briefly mentioned.
- ✓ Relationship between different risk factors, type of diabetes and symptoms has been depicted.
- ✓ The effect of diabetes on health, wealth and life has been presented.
- ✓ Different data mining and soft computing techniques have also been explored.
- ✓ Details of research articles published in last 10 years have been analyzed.
- ✓ Article inclusion and exclusion criterion is mentioned.
- ✓ Extensive analysis of different mining techniques has been presented.
- ✓ Effect of pre-processing techniques on the predictive rate of accuracy has been shown.
- ✓ Rate of accuracies achieved using individual and hybrid approaches are also presented.
- ✓ Diabetes diagnosis using image data.

Methodology: Data mining and soft computing techniques are effectively used for disease diagnosis. The remaining part of this section will explain data mining and soft computing techniques.

Data Mining Techniques: With the explosive database growth, information retrieval is not enough for decision making. The complete analysis of both data and information must be reviewed to take better decisions.

Data Mining aims at discovering knowledge from the data and represents it in the easily understandable form for humans. Data mining is collection of techniques needs to extract and exhibit large data to determine hidden predictive information.

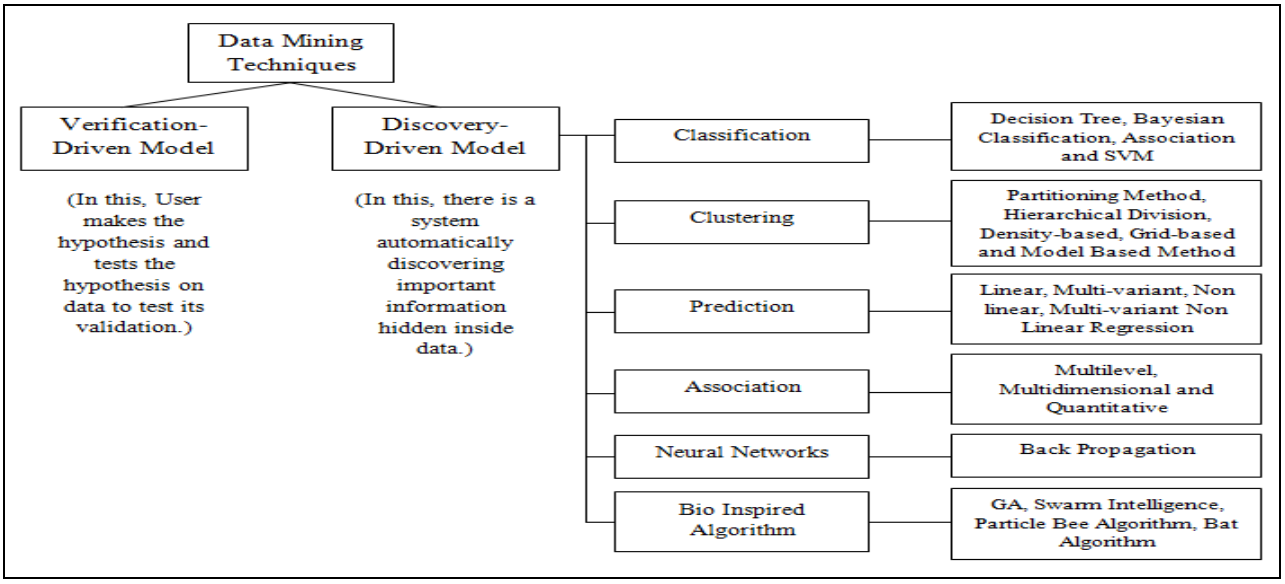

FIG. 4: CLASSIFICATION OF DATA MINING TECHNIQUES

In general, data mining techniques can be classified as verification and discovery techniques. In the verification-driven technique, initially, users formulate the hypothesis and then verify the data to check its validity. This technique performs the operations like querying, reporting, validating hypothesis, statistical and multidimensional analysis. Verification - driven model emphasis on user whereas discovery-driven model emphasis on system. These systems mined the databases automatically to discover information that is concealed in the data. Discovery-driven techniques are focused on classification, clustering, prediction and association rule mining. Classification is a process of finding rules to assign new objects into a predefined category.

Classification can be performed in two steps. In first step classifier is build by analyzing training set and class label attributes. In second step a model is used for classification of test set (independent of training set). Some of the important classifiers are decision tree, naive bayes, SVM, heuristic algorithms, neural networks etc. Decision tree is like a flow chart that classifies instances depending upon on the features. Each internal node represents the test case, branches show results of tests and leaf nodes hold the labels of classes. This technique performs better when there are discrete features. Some algorithms used for inducing decision tree are Classification and Regression Tree (CART), ID3, C4.5, and Chi-squared Automatic Interaction Detection (CHAID). CART is a binary decision tree algorithm that follows the pattern of structured queries to determine their answers.

The algorithm initialize with the rules to dole the data, stopping rules to introduce terminals and finally predicting target variable. To find optimal solution for classification, ID3 is used with the minimized depth of decision tree. In ID3, data is sorted to get the best split at every node whereas in C4.5, one attribute is selected to split the samples into subsets. Bayesian classification is based on bayes theorem that is used to solve the diagnostic and predictive problems. Bayesian classifiers are statistical classifiers also called naive bayesian classifier. Naive bayes algorithm is applied on very large datasets. These are based on conditional probabilities and are easy to develop. This algorithm finds the probability of occurring events by using the probability of already occurred events. Association rules are used to realize the interesting patterns or correlations among the objects in the large data sets. In addition, SVM is used for classifying both linear and non-linear data. It incorporates the structured risk minimization to decrease the error and to improve the effectiveness of classification. SVM classifier use data points to create hyper plane and maximize the difference between data points by using SVM [23, 24].

Clustering means grouping of physical objects into the set of similar object class (cluster). It depends on the concept of distance or similarity metrics. It is also called data segmentation as it partitions the large datasets into the similar groups. The various methods used for clustering are partitioning, hierarchical, density-based, grid-based and model-based method. In partitioning method, one has to first declare the number of clusters before making groups. This method is used to cluster large datasets with complex shapes by using K-means and K-mediods approaches. Hierarchical method creates the hierarchical decomposition of dataset either by using top-down (called divisive) approach or bottom-up (called agglomerative). Density-based method discovers the clusters of arbitrary shapes based on the density connectivity and distribution value analysis of density function. The three approaches used in this method are Density Based Spatial Clustering of Applications with Noise (DBSCAN), Ordering Points to Identify the Clustering Structure (OPTICS) and Density-Based Clustering with Constraints (DENCLUE). Grid-based methods form the clusters from continuous group of dense cells. It eliminates the cells whose density is below the threshold. Generally grid based techniques deals with independent number of objects and expedites the clustering process. STING and Clustering in Quest (CLIQUE) algorithms are used in the method. In model-based method, for each cluster a model is hypothesized to get the best fit data on given model. COBWEB and self organizing feature map (SOM) are two important model based methods [23].

Prediction is a special case of classification in which output is a future value, usually a single value, either categorical or numerical. It is one of the imperative data mining tasks based on two types of variables *viz.* dependent and independent.

The major objective of prediction is to attain the value of dependent variable based upon independent variables. Some common methods required for forecasting are regression, time series, Principle Component Analysis (PCA), SVM, GA etc. Association rules assist in finding a connection between two or more data items. Market basket analysis and apriori algorithm are best examples of association discovery [24].

The performance of early diagnostic healthcare system is based upon the rate of accuracy achieved in predicting the disease. Generally, different measures related to predictive accuracy are represented in the form of confusion matrix (2 × 2). It represents from attributes True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Here, TP and TN represent the number of positive and negative correctly classified events. Similarly FP and FN represent the positive and negative incorrectly classified events by the diagnostic system [24]. In addition, accuracy, sensitivity, sensitivity, positive precision and error rate are some other important measures of predictive system and are represented in **Table 1**.

TABLE 1: CONFUSION MATRIX FOR CLASSIFICATION PROBLEM

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

Accuracy is the ratio of truly classified events to the total events, mathematically:

$$\text{Accuracy} = (TP + TN) -/(TP + TN + FP + FN) \quad ...(1)$$

Sensitivity is also called TN rate. It shows that the correctly classified negative cases can be determined by:

$$\text{Sensitivity} = TP / (TP + FN) \quad ... (2)$$

Specificity is also called TP rate that shows the correctly identified positive cases, can be determined by:

$$\text{Specificity} = TN / (TN + FP) \quad ... (3)$$

Precision is the proportion of correctly predicted positive cases and can be determined by:

$$\text{Positive Precision} = TP / (TP + FP) \quad ... (4)$$

Error rate is also called misclassification rate that shows incorrect classified cases, can be determined by:

$$\text{Error Rate} = (FP + FN)/(TP + TN + FP + FN) \quad ... (5)$$

Soft Computing Techniques: Soft Computing encourages the integration of methodologies that aims to easily design the solutions of real life problems that are difficult to model. Soft computing techniques are the blending of distinct methodologies that were designed to solve multifaceted real World problems (medical science, management, agricultures, economics *etc.*) that were intractable to solve otherwise [25]. The major features of soft computing techniques are uncertainty, imprecision and approximation tolerance. It has been extensively premeditated and applied for engineering and healthcare computing. Generally, soft computing techniques are categorized as exact and approximate optimization techniques. **Fig. 5** represents the subcategories of these techniques.

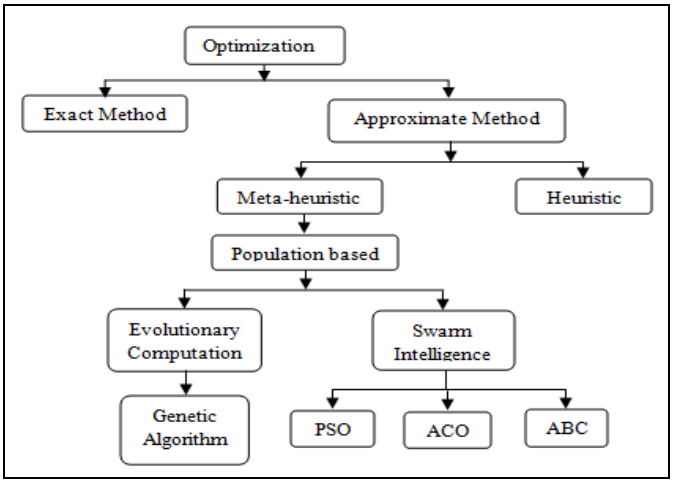

FIG. 5: CLASSIFICATION OF COMMON OPTIMIZATION TECHNIQUES

Exact method requires momentous computation time. Therefore, are not recommended for real world problems. Additionally, approximate methods are decomposed as heuristic and meta-heuristic. Heuristic methods can be practically implemented to get high quality solutions with less computation time. Meta-heuristics methods can be applied on large and complex real world problems to find the optimal solutions. Neural Networks (NN) are inspired by the biological system to solve the problem using complex interconnection of simple processing elements.

Neural networks are information driven rather than data driven. The most common networks used in NN is the Back Propagation Network (BPN) that consists of an input layer, and an output layer with one or more intermediate hidden layers [26].

Genetic Algorithm commonly abbreviated as 'GA'. It is a prominent evolutionary approach. The general conception of 'Genetic Algorithm' was proposed by John Holland [27]. These are search algorithms explicitly intended to imitate the principle of the natural biological evolution process. 'GA' borrows its indispensable features from natural genetics. It is a stochastic technique that lay down good-quality solution with low time complexity. It sanctions a population composed of many individual chromosomes to evolve under delineated selection rules to breed a state that optimize the objective function. GA lucratively operates on a population of solutions rather than a single solution. It generally employs some heuristics like 'Selection', 'Crossover' and 'Mutation' to develop better solutions [28].

Evolutionary and swarm intelligence are the major population based meta-heuristic techniques. Swarm Intelligence is an innovative optimization technique inspired by the biological behaviour of animals, birds and fish. The inspiration often comes from the swarming, flocking and herding phenomena in vertebrates. Swarms are the individual agents who interact with one another and with their environment [29]. They follow the simple rule of not having the centralized control structure, and leads to the emergence of complex global behaviour. It is basically used to apply the self organizing behaviour of the mechanical agents through nearest neighbour interaction. The common swarm intelligence algorithms are ant colony optimization (ACO) and particle swarm optimization (PSO). The use of swarm intelligence has been extended from conventional optimization problems to optimization -based data mining to solve discrete and continuous optimization problems.

Particle Swarm Optimization is an adaptive algorithm introduced by Dr. Russell C. Eberhart and Dr. James Kennedy in 1995. It is a meta-heuristic, population-based stochastic optimization technique inspired by the group behaviour of organisms [30]. It is used to search the optimal solution in the search space. The algorithm consist of two operators *i.e.* velocity and position update. Algorithm starts with the random initialization of particles. Each particle has a position in search space (piD), where D is dimension. The velocity of particle is represented by (viD). Each particle is accelerated toward the particles previous best position (pbest) and the global best position (gbest). In each iteration, new velocity and position of the particle is calculated [27].

Artificial Bee Colony (ABC) algorithm was introduced by Dervis Karaboga in 2005, which was motivated by the foraging behaviour of honey bees. The aim of bees is to locate the places of food sources with highest nectar amount [31]. The colony of bees categorized into three working patterns; the bees who are currently exploiting the food sources are employed bees, the bees those are waiting on the hive are onlookers and those who leave the hive to discover new food sources are scouts. The employer bees share the information of food sources with the onlookers by using a technique called waggle dance. Once the food collected from a source it becomes useless and scout bees start searching the new source. Food source positions are the possible solutions of the problem. Amount of nectar in food represents the quality of solution [27].

Research Publication and Article Selection Criteria: Following queries have been fired to determine the status of number of articles published related to diabetes diagnosis using data mining and soft computing techniques.

- ✓ "Diabetes Diagnosis" using naive bayes
- ✓ "Diabetes Diagnosis" using C4.5 algorithm
- ✓ "Diabetes Diagnosis" using ID3 algorithm
- ✓ "Diabetes Diagnosis" using random forest
- ✓ "Diabetes Diagnosis" using SVM
- ✓ "Diabetes Diagnosis" using KNN
- ✓ "Diabetes Diagnosis" using ACO
- ✓ "Diabetes Diagnosis" using PSO
- ✓ "Diabetes Diagnosis" using Artificial Bee Colony
- ✓ "Diabetes Diagnosis" using Genetic algorithms

Here, **Table 2** represents number of articles published (2008 to 2017) and indexed in Google scholar database.

TABLE 2: COMPARISON OF RESEARCH DONE IN LAST TEN YEARS WITH DIFFERENT CLASSIFICATION TECHNIQUES

Year	Classification Techniques									
	Naive Bayes	C4.5	ID3	Random forest	SVM	KNN	ACO	PSO	ABC	GA
2008	9	8	1	28	16	7	1	2	0	176
2009	7	6	1	21	17	3	3	9	0	89
2010	13	10	1	27	29	13	1	7	0	114
2011	8	15	5	31	34	9	9	4	0	127
2012	20	29	10	53	46	16	14	14	0	162
2013	20	34	10	63	47	17	8	9	3	224
2014	50	45	13	68	92	40	29	25	6	250
2015	62	56	16	96	126	55	28	24	9	289
2016	79	80	26	120	135	71	38	44	14	308
2017	103	108	26	157	173	90	48	52	26	440

From **Table 2**, it is observed that in last ten years, ample of research has been done using different data mining and soft computing techniques for diagnose diabetes. As compared to year 2008, in 2017 usage of C4.5, KNN, naive bayes, and SVM have been increased by 13, 13, 11 and 11 times respectively. Similarly, rate of usage of ID3, ACO, PSO and ABC have boost up by 26, 44, 26 and 26 times respectively. In last ten years, ABC has been least used for diabetes diagnosis.

GA constantly witnessed large number of article publications right from 2008 to 2017 as compared to naive bayes, C4.5, random forest, ID3, KNN, SVM, ACO, PSO, and ABC.

Here, the articles published in 2010 and onwards are considered in the review. There is 3622 number of articles published in the period. After successful screening (four level), 63 articles are finally incorporated in this study.

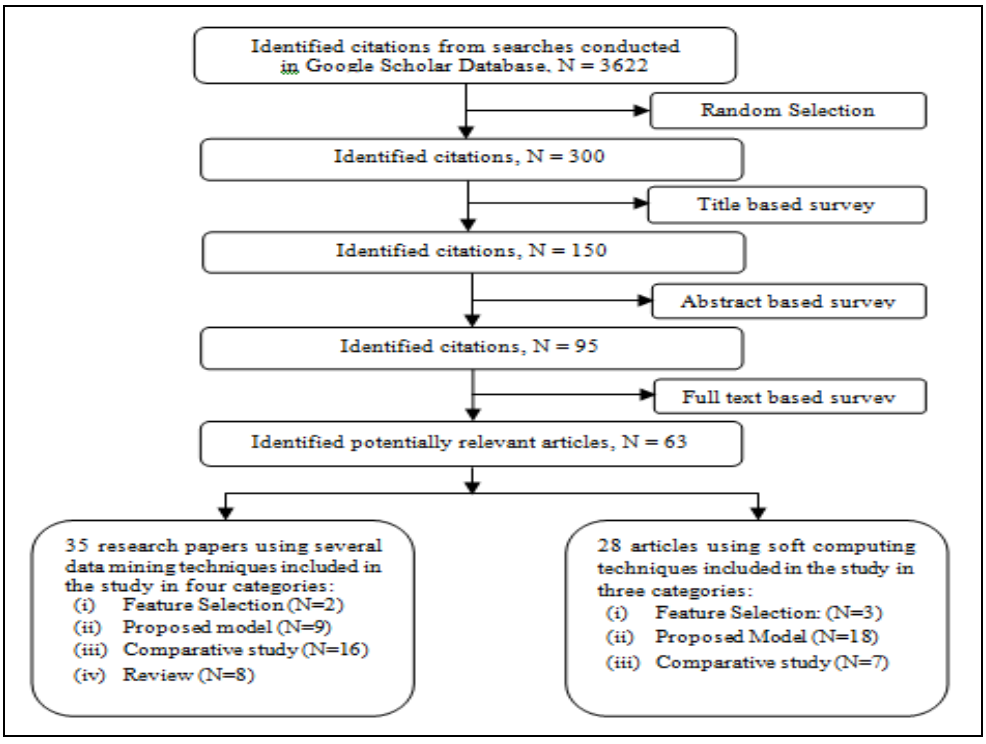

FIG. 6: ARTICLE SELECTION CRITERION

According to Google scholar database, it is observed that the most of the work has been done on gestational and type 2 diabetes diagnosis (*i.e.* 44% and 37% respectively). And little attention has been paid on diagnosing type1 diabetes. Whereas, no research have been done for diagnosing type 3 and type 4 diabetes using data mining and soft computing techniques. **Fig. 7** represents the number of research articles indexed on different types of diabetes using data mining and soft computing techniques.

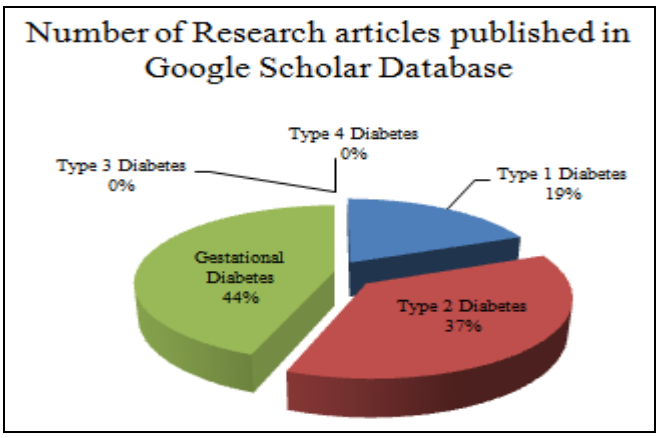

FIG. 7: PERCENTAGE OF DIABETES DIAGNOSIS BASED ARTICLES INDEXED IN GOOGLE SCHOLAR

RESULTS AND DISCUSSIONS:

Dataset: The early and affective diagnosis of diabetes is heavily based upon the datasets of the patients. Generally, dataset represents the number of attributes, instances and feature set used while diagnosis of diabetic patients. **Table 3** represents some common feature set used in diabetes diagnosis with data type.

For prospecting diabetes in humans, author generally used datasets University of California, Irvine Machine Learning Repository (UCIMLR), Pima Indian Diabetes Database (PIDD), SAS, World Health Organization (WHO) and different hospitals, clinics and health centres. The range of attributes in these datasets lies between 6 to 63. However, number of instances lies between 80 and 50,000. **Table 4** represents the year wise details of some of the key authors.

TABLE 3: REPRESENTATION OF COMMON FEATURE SET WITH DATA TYPE

Feature set	Range	Data type
Age	In (years)	Integer
Gender	Male, Female	Discrete
Family History (Diabetic)	Yes, No	Discrete
Number of times pregnant	>= 0	Integer
Plasma Glucose concentration	In (mg/dl)	Real
Diastolic Blood Pressure	In (mm Hg)	Real
Body Mass Index	In (Kg m^2)	Real
Triceps skin thickness	In (mm)	Real
Two hour Serum insulin	In (mu U/ml)	Real
Frequent Urination	Yes, No	Discrete
Thirst	Yes, No	Discrete
Hunger	Yes, No	Discrete
Physical exercise	Yes, No	Discrete
Smoke	Yes, No	Discrete
Alcohol	Yes, No	Discrete
Tongue color	-	Image
Tongue texture	-	Image
Iris region	-	Image
Iris pupil	-	Image
Iris sclera	-	Image

TABLE 4: REPRESENTATION OF DATASET WITH COMMON FEATURES SET

Author	Year	Dataset	Attributes	Instances	Data Type
Pardha Repalli [32]	2011	SAS	37	50000	Integer, discrete, text, real
A. Aljumah *et al.*, [33]	2012	WHO	5	Not mentioned	Integer, discrete
F. Beloufa *et al.*, [57]	2013	UCIMLR	9	760	Integer, discrete, real
Nagarajan *et al.*, [36]	2014	Different Hospitals and clinics	6	600	Integer, discrete
P. Radha *et al.*, [39]	2014	UCIMLR	9	760	Integer, discrete, real
Sankaranarayanan *et al.*, [40]	2014	PIDD	9	770	Integer, discrete, real
Varma *et al.*, [71]	2014	UCIMLR	9	336	Integer, discrete, real
H. R. Sahebi *et al.*, [63]	2015	UCIMLR	8	760	Integer, discrete, real
A. Pavate *et al.*, [67]	2015	Not mentioned	15	250	Integer, discrete, text, real
A. Bansal *et al.*, [72]	2015	Not mentioned	Not mentioned	80	Image
D. K. Choubey *et al.*, [69]	2016	UCIMLR	8	760	Integer, discrete, real
J. Zhang *et al.*, [73]	2016	TCM Health Centre, Shanghai	23	827	Image
R. Asgarnezhad *et al.*, [74]	2017	UCIMLR	8	760	Integer, discrete, real
P. Samant *et al.*, [75]	2018	Not mentioned	63	200	Image

Diabetic Diagnosis using Data Mining Techniques: Pardha Repalli (2011) applied the statistical techniques to predict diabetes. After comparing decision tree and regression, authors found that decision tree has given the highest classification rate (0.50) and lowest error rate (0.043). Author concluded that the risk of diabetes increases with the increase in the age of individual [32]. Abdulllah A. Aljumah *et al.*, (2013) used regression based data mining technique over two different age groups (young and old). The study used to predict the preferential order of treatments according to patient's age in Saudi Arabia.

Predictions were forecasted using different metrics like drug, diet, weight, smoke cessation, exercise and insulin. Author concluded that the drug treatment in younger can be delayed but in old age it must be prescribed immediately [33]. K. Rajesh *et al*., (2012) used various classification methods *viz*. C-RT, CS-RT, C4.5, ID3, KNN, Linear Discriminant Analysis (LDA), naive bayes, Partial Least squares- Discriminant Analysis (PLS-DA), SVM, RND tree for diabetes diagnosis. Among all the methods, C4.5 has shown the best accuracy (~91%) [34]. K. R. Lakshmi *et al*., (2013) performed the comparison between ten data mining algorithms- C4.5, SVM, KNN, Prototype Neural Network (PNN), Bayesian Logistic Regression (BLR), Multinomial Logistic Regression (MLR), PLS-DA, Partial Least squares-Linear Discriminant Analysis (PLS-LDA), k-means and apriori to get best results for diabetes diagnosis in patients. Authors found that PLS-DA gives optimal results as compared to other techniques. The rate of accuracy achieved with PLS-DA was 74% [35].

Srideivanai Nagarajan *et al*., (2014) used four algorithms ID3, naive bayes, C4.5 and random tree to diagnose gestational diabetes in pregnant women. Authors found that random tree give better predictive results for gestational diabetes. Rate of accuracy achieved using random tree is 93.8% [36]. In order to predict diabetes mellitus with the better accuracy, Veena Vijayan *et al*., (2014) discussed five data mining algorithms – Expectation-Maximization (EM) algorithm, KNN algorithm, K-means algorithm, Amalgam KNN algorithm and Adaptive Neuro Fuzzy Inference System (ANFIS). Among all algorithms, ANFIS with adaptive KNN provide better classification rate (80%) [37]. Thangaraju *et al*., (2014) diagnose diabetes by implementing best first search technique and greedy forward selection method. The goal of this paper was to predict the diabetes type in a patient of different age groups with different lifestyle habits [38]. In context to diabetes diagnosis system, P. Radha *et al*., (2014) compared five different supervised classification algorithms namely C4.5, SVM, KNN, PNN, and BLR.

The study concluded that BLR diagnose diabetes with best accuracy (75%) and least error rate (0.27) [39]. Sankaranarayanan *et al*., (2014) used two classification techniques *viz*. FP-Growth and apriori for diabetes diagnosis. The number of rules generated by both algorithms were same and helps to enhance the classification performance [40]. M. Mounika *et al*., (2015) used three classification algorithms ZeroR, OneR and naive bayes to predict type of diabetes. Among all the three algorithms naive bayes has given best results with precision 0.975, TP rate 0.974, correctly classified instances 97.43% and incorrectly classified instances 2.56% [41]. N.M. Saravana Kumar *et al*., (2015) used predictive analysis algorithm in Hadoop to diagnose diabetes effectively in the patients of rural areas. The study also used to find the availability of treatment with reduced cost in remote locations for diabetic patients [42].

Komal Agicha *et al*., (2015) performed predictive analysis to diagnose diabetes in patients by considering the parameters like age and gender. Authors have used two classification techniques *viz*. naive bayes and decision tree with R studio tool in the study to predict the disease and help healthcare professionals to suggest the effective treatment [43]. Aiswarya Iyer *et al*., (2015) used naive bayes and J48 decision algorithm with Weka tool to predict diabetes quickly and efficiently in human. The study also helps health professionals to predict the effective treatment for the patients with different age groups [44]. Sajida Perveen *et al*., (2016), illustrated risk factors of diabetes over the patients with different age groups. The performance of Adaboost ensemble was stupendous [46]. T. Daghistani *et al*., (2016) compared three data mining algorithms *viz*. SOM, C4.5 and random forest. The study used to diagnose diabetes in the patients of Saudi Arabia. Authors used 66325 samples with 18 features set. The study concluded that the random forest has shown best precision with 68% whereas SOM and C4.5 have shown 46% and 67% respectively to predict diabetes in patients [47]. M. Teimouri *et al*., (2016) conducted a study for different genetic and life style based diseases *viz.* Hypothyroidism, hyperthyroidism, migraine, anaemia, infection, hypertension, diabetes, asthma, epilepsy, seizure, osteoporosis and calcium deficiency. 1412 different samples were analyzed using six different classification algorithms *viz*. decision tree, SVM, NN, naive bayes, Logistic regression (LR) and KNN where SVM has shown the better accuracy with 95.32% [48].

TABLE 5: REVIEW ON ISSUES AND OUTCOME TO PREDICT DIABETES WITH VARIOUS DATA MINING TECHNIQUES

Author (year)	Diabetes Type	Issue	Data Mining Technique	Software	Outcome
A. A. Aljumah *et al*., [33] (2012)	Type 1 and 2	Risks in different age groups	Regression based technique and SVM algorithm	Oracle Data Miner 10.1	Predict high risk of diabetes in old aged people
K. Rajesh *et al*., [34] (2012)	Type 1 and 2	To find best classification method for predicting state of diabetes	C-RT, CS-RT, C4.5, ID3,KNN, LDA, Naive bayes, PLS-DA, SVM, RND tree	Not mentioned	C4.5 outperforms other techniques with 91% of accuracy
K. R. Lakshmi *et al*., [35] (2013)	Type 1 and 2	Predicting diabetes with best accuracy	C4.5, SVM, KNN, PNN, BLR, MLR, PLS-DA, PLS-LDA, k-means, Apriori	Tanagra tool	PLS-DA algorithm was best with less computing time, error rate and high accuracy
S. Nagarajan *et al*., [36] (2014)	Gestational and Type2 Diabetes	To study nature of diabetes in pregnant women	ID3, Naive bayes, C4.5 and Random tree	WEKA and Tanagra	To predict the risk of type 2 diabetes in gestational mother and children, random forest seems to give optimal results. The rate of accuracy achieved using random forest for type 2 diagnosis in pregnant women is 93.84%
V. Vijayan *et al*., [37] (2014)	Type 1 and 2, Gestational and Congenital Diabetes	Predict the disease with best accuracy	EM algorithm, KNN, K-Means, Amalgam KNN and ANFIS	Not mentioned	ANFIS has shown better accuracy for forecasting diabetes
P. Thangaraju *et al*., [38] (2014)	Type 1 and 2	Predict factors responsible for diabetes in the patients with different age groups	Best first search and Greedy forward selection method	Not mentioned	One is able to get instant and precise diagnosis using Best first search and Greedy forward selection method
P. Radha *et al*., [39] (2014)	Type 1, Type 2 and Gestational diabetes	Diabetes diagnosis with different class-ification techniques	C4.5, SVM, k-NN, PNN, and BLR	Tanagra	BLR has given better results as compared to C4.5, SVM, KNN and PNN
M. Mounika *et al*., [41] (2015)	Type 1 and 2	Comparative analysis for diabetes diagnosis	ZeroR, OneR and Naive Bayes	WEKA	Naive bayes algorithm was fastest and ZeroR is slowest to diagnose diabetes
K. Agicha *et al*., [43] (2015)	Type 1 and 2	Aged based analysis of diabetes diagnosis	Decision tree, Bayesian network algorithm	R studio	Predict effective treatments of diabetes for patients with different age group
A. Iyer *et al*., [44] (2015)	Type 1, Type 2 and Gestational diabetes	Diagnose diabetes by analysing patterns	J48 decision tree and Naive bayes algorithm	WEKA	Decision tree and naive bayes predict the disease quickly and effectively
P. C. Thirumal *et al*., [49] (2015)	Type 1 and 2	Predict disease with highest accuracy	Naive bayes, C4.5, SVM, KNN	WEKA	C4.5 diagnose diabetes with highest accuracy and SVM with lowest accuracy
S. Sa'di *et al*., [50] (2015)	Type 2	Predict diabetes in patients with highest accuracy	Naive bayes, Radial Basis Function (RBF) network and J48	WEKA	Naive bayes algorithm has shown the highest accuracy for diabetes diagnoses
S. Perveena *et al*., [46] (2016)	Type 1 and 2	Risk factors of diabetes in patients with different age gr	Adaboost ensemble, bagging and standalone J48 decision tree algorithm	WEKA	Old aged people are more prone to diabetes as compared to young ones
T.P Kamble *et al*., [51] (2016)	Type 1 and 2	Predict the type of diabetes effectively	Deep learning based restricted boltzmann machine approach and J48 decision tree algorithm	Not mentioned	Deep learning approach classify the patient is diabetic or not and decision tree was used to predict type of diabetes
S. Hina *et al*., [52] (2017)	Type 2	Predict the type of diabetes effectively	Naive bayes, MLP, ZeroR, J.48, Random Forest, and LR	WEKA	MLP outperforms other techniques with 81.8% of accuracy
S. Selvakumar *et al*., [53] (2017)	Type 1 and 2	Diagnose diabetes with highest accuracy	Binary LR, Multilayer Perceptron and K-Nearest Neighbour	Not mentioned	The rate of accuracy achieved with KNN is 80%
Pawar *et al*.,[19] (2017)	Type 2	Diagnose diabetes with different classification techniques	SVM, AI, ANFIS and KNN, and PCA , CART	Not mentioned	SVM was better among all for diabetic diagnosis

T.P Kamble *et al*., (2016) proposed a deep learning based restricted boltzmann method approach with J48 decision tree algorithm to predict risk of diabetes in patients. Restricted Boltzmann method was used to effectively classify dataset and J48 algorithm was used to detect type of diabetes in patient. The rate of precision achieved by using this proposed model was 77.8% [51].

The best rate of diabetes diagnosis accuracy achieved using data mining techniques by different authors is depicted in **Fig. 8**. Based upon data and feature set, the predictive accuracy lies between 68 to 96%.

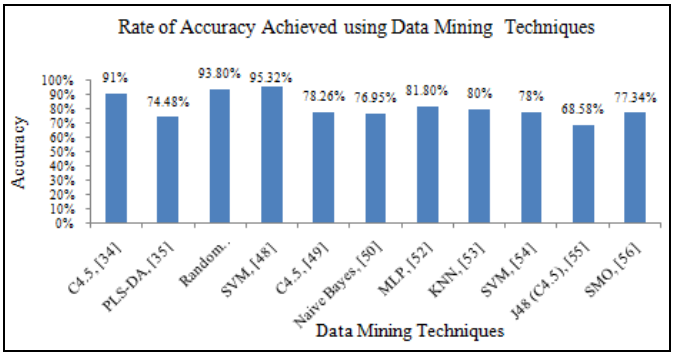

FIG. 8: RATE OF ACCURACIES ACHIEVED USING DATA MINING TECHNIQUES

Diabetic Diagnosis using Soft Computing Techniques: Dario Gregori *et al*., (2011) showed that for diabetes diagnosis, simple model outperforms the complex models. The study was conducted with the data mining statistical techniques where the six classifiers viz. LR, GAM, aPPR, LDA, QDA, ANN were applied. The study concluded that GAM which was a simpler algorithm and shows better predictive accuracy as compared to ANN [57]. Mythili Thirugnanamet *et al*., (2012) presented the two stage approach to predict diabetes. The first stage was prediction stage used three techniques fuzzy logic (F), neural network (N) and case based reasoning (C), known as FCN approach. Second stage applied rule based algorithms on the outcomes from first stage. The study concluded that the proposed method gives the high classification rate to predict diabetes [58]. Jefri Junifer Pangaribuan *et al*., (2014) conducted the study to diagnose diabetes in women aged 21 years. ELM method of ANN was used to forecast the disease. The results of ELM were compared with back propagation method to know the accuracy and less error rate.

The study has shown that error rate of ELM and back propagation are 0.4036 and 0.9425 respectively [59] Turker Ince *et al*., (2010) proposed the assessment method for the medical diagnosis of diseases like breast cancer, diabetes and heart diseases. The study compared the performance of two conventional techniques back propagation and PSO. Authors found that classification rate witnessed with PSO is better [60]. Dervis Karaboga *et al*., (2010) have shown that the performance of ABC was better than fuzzy clustering. The study concluded that PSO achieved lowest Mean Squared Error (MSE) levels. However, back propagation was more efficient than PSO [61].

Sapna *et al*., (2012) applied combination of GA and fuzzy system to predict Type 2 diabetes in patients. The study has shown that the fuzzy genetic algorithm was more efficient than the classical GA [62]. Mustafa Serter Uzer *et al*., (2013) applied a hybrid approach of ABC algorithm and SVM to predict disease. ABC algorithm was used for feature selection and SVM was used for classification. The study shows the classification accuracy with 79.29% for diabetes diagnosis[63]. Fayssal Beloufa *et al*., (2013) proposed modified version of ABC algorithm to predict the disease with better accuracy. The solution quality achieved with modified ABC algorithm is 3% better than the basic ABC algorithm [64]. Kovalan *et al*., (2014) used genetic algorithm to analyze large dataset and classify diabetes in human. The study provided diagnosis specification of diabetes with less time consequences over large databases [65].

K. Vijaya Lakshmi *et al*., (2014) proposed a model to predict diabetes by analysing gene and chromosomes using GA. The proposed model also assists to predict the disease by analysing the general symptoms in patients [66]. Omar S. Soliman *et al*., (2014) proposed a model for diagnosing type 2 diabetes in patients. The proposed algorithm used LS-SVM to perform classification and Modified-PSO to perform optimization of data. With this approach authors got good (97.8%) predictive rate of accuracy over 800 patients [67]. S. Karthikeyan *et al*., (2014) designed a meta-heuristic technique Particle Swarm Artificial Bee Colony (PSABC) to diagnose cancer and diabetes. Authors combine features of both PSO and ABC. The intension of this model was to reduce the average error rate.

Authors showed that by using PSABC, they were able to reduce the average error rate by 5.35% and 5.33% as compared to the average error found with individual approach of PSO and ABC [68].

To diagnose diabetes disorder E. Sreedevi *et al*., (2015) proposed Threshold Genetic Algorithm (TGA). TGA is combination of simple GA and Minkowski distance method. In TGA, fitness function is formulated using minkowski distance method. TGA gives better predictive accuracy as compared to simple GA. Rate of accuracy achieved using TGA is 72.2% [69]. Hamid Reza Sahebi *et al*., (2015) proposed a fuzzy classifier by using a modified PSO and fuzzy logic for early diabetes diagnosis. The study compared the performance of basic PSO and modified PSO. The rate of predictive accuracy achieved using basic and modified PSO are 78.5% and 85.1% respectively [70].

For early diabetes diagnosis, C. V. Subbulakshmi *et al*., (2015) proposed a hybrid meta-heuristic swarm intelligence technique. Authors combine the features of PSO and ELM. PSO assist in parameter selection. ELM helps in enhance the performance of PSO. It was observed that ELM with self regulating PSO gives better performance than simple PSO [71]. Liwei Zhang *et al*., (2015) presented Kernel based Extreme Learning Machine (KELM) with PSO for diagnosis diabetes. The study has shown KELM performed better than grid algorithm and diagnose the disease more accurately than SVM [72]. Omar S. Soliman *et al*., (2015) proposed an algorithm using bat algorithm with chaotic levy flight that assist in early diagnosis of diabetes. The study proved that this proposed algorithm was better than the traditional bat algorithm [73]. Aruna Pavate *et al*., (2015) devised a hybrid meta-heuristic model for diabetes diagnosis. Authors combine the features of GA, KNN and fuzzy system. Authors used GA and KNN for diabetes prediction. Additionally, fuzzy rules were incorporated to analyze the other risks associated with diabetes. The rate of accuracy, specificity and sensitivity achieved using hybrid meta-heuristic model are 95.50%, 86.95% and 95.83% respectively [74].

E. Sreedevi *et al*., (2016) proposed a Hybrid Genetic Classifier Method (HGCM) for diagnosing type 2 diabetes in patients. Authors combine GA with KNN distance methods *viz*. manhattan, euclidean, chebyshev and minkowski distance methods. GA was used to classify the dataset and Distance methods used as fitness function. It was observed that Minkowski distance method give better accuracy [75]. Dilip Kumar Choubey *et al*., (2016) proposed a model by using GA and Multilayer Perceptron Neural Network (MLP NN) to predict diabetes in human. Feature selection was performed on dataset of 768 patients by GA and classification on the dataset was performed by MLP NN.

The predictive rate of accuracy of proposed model was 79.13% [76]. M. Komi *et al*., (2017) have compared the performance of five different algorithms *viz.* GMM, ANN, ELM, LR and SVM, for early prediction of diabetes. The study has shown ANN performed better the other four algorithms and achieved 89% as highest predictive rate of accuracy [77].

TABLE 6: REVIEW ON ISSUES AND OUTCOME TO PREDICT DIABETES WITH VARIOUS SOFT COMPUTING TECHNIQUES

Author (year)	Diabetes Type	Issue	Data Mining Technique	Software	Outcome
E.P. Ephzibah [78] (2011)	Type1 and 2	Diagnose diabetes with reduced cost	Fuzzy logic Based and GA	MATLAB	Fuzzy logic with GA diagnose diabetes with accuracy 87% where as fuzzy logic classifier predict the disease with 68% accuracy
M. S. Uzer *et al*., [63] (2013)	Type1 and 2	Pre processing and precise diagnosis	ABC and SVM	Not mentioned	ABC algorithm was used to remove redundant feature and SVM used to classify features to predict liver disorder and diabetes with accuracy 74.81% and 79.29% respectively
F. Beloufa *et al*., [64] (2013)	Type1 and 2	To devise better version of ABC for diabetes diagnosis	ABC and fuzzy method	Not mentioned	Modified ABC with fuzzy method act as a powerful tool for diabetes diagnosis. The precision rate accomplished using modified ABC is 84.25%
Divya	Type1 and	Predict risk of	Artificial Neural	Not	ANN with back propagation method

et al., [79] (2013)	2	diabetes and its type in patients	Networks (ANN) and Back propagation method	mentioned	predict diabetes accurately
S. Kovalan *et al*., [65] (2014)	Type1 and 2	To predict diabetes with high accuracy in less time	GA with Roulette wheel selection	Not mentioned	GA assists in expediting the diagnosis process for large sized database
Varma *et al*., [80] (2014)	Type1 and 2	Early diagnosis of diabetes	Rough Set Theory	MATLAB	Predict diabetes early and effectively with 77.9% classification rate
S. Aishwarya *et al*., [81] (2014)	Type1 and 2	Diagnose diabetes with highest accuracy	GA and Least Squares-Support Vector Machine (LS-SVM)	Not mentioned	GA and LS-SVM algorithm has given highest accuracy than SVM, DSS, ACO-SVM and grid algorithm
J. J. Pangaribuan *et al*., [59] (2014)	Type1 and 2	Diagnose diabetes using Extreme Learning Machine (ELM) with accuracy than back propagation	ELM with ANN, back propagation	MATLAB	ELM achieved better accuracy and less error rate than back propagation
E. Sreedevi *et al*., [69] (2015)	Type1 and 2	Diagnose diabetes using GA and Minkowski distance method	GA, Minkowski distance metrics	Not mentioned	TGA diagnose diabetes with better accuracy (72.214%) than GA
A. Pavate *et al*., [74] (2015)	Type1 and 2	Predict diabetes and complication associated with it	GA, KNN, Fuzzy-rule based system	Not mentioned	A GA based KNN technique assist to forecast three year risk for diabetes
T. Santhanam *et al*., [82] (2015)	Type1 and 2	To design a meta-heuristic based diabetes diagnosis model	K-mean, GA and SVM	Not mentioned	Authors achieved optimal rate of predictive accuracy. The accuracy rate achieved with hybrid meta-heuristic model is 98.82%
D. K. Choubey *et al*., [83] (2015)	Type1 and 2	Diabetes diagnosis with least cost and improves ROC	GA _J48graft decision tree	Not mentioned	GA_J48graft decision tree successfully diagnose diabetes. The hybrid use of GA and J48graft improves predictive accuracy of J48graft by 4.4%
E. Sreedevi *et al*., [75] (2016)	Type2	Diagnose diabetes using hybrid GA	HGCM, Minkowski distance metrics	Not mentioned	HGCM predict diabetes accurately
D. K. Choubey *et al*., [76] (2016)	Type1 and 2	Improve accuracy of classification with reduced computation cost and time	GA, MLP NN	Not mentioned	GA used for feature selection and MLP NN used to classify features for accurate diagnoses with less computation cost and time
R. Asgarnezhad *et al*., [84] (2017)	Type 2	Efficiency of pre-processing techniques on diabetes dataset	Forward selection and backward elimination, brute force, GA and SVM	Not mentioned	GA is used as pre-processing technique to handle missing value and optimize selection. SVM is used as classifier to diagnose diabetes mellitus. The predictive rate of accuracy was 84.35%
Rashmi *et al*., [85] (2017)	Type1 and 2	Diabetes diagnosis with highest accuracy	ACO and SVM	MATLAB	Diagnosis of diabetes with hybrid method using ACO and SVM has shown accuracy 99.2%
K. Ateeq *et al*., [86] (2017)	Type1 and 2	Optimization of diabetes dataset using PSO to improve diagnosis	PSO and ANN	Not mentioned	ANN with PSO diagnose diabetes with accuracy 82.4% where as ANN classifier predict the disease with 75.3% accuracy
M. Komi *et al*., [77] (2017)	Type1 and 2	Prediction of diabetes with best accuracy rate	Gaussian Mixed Model (GMM), ANN, ELM, LR and SVM	Not mentioned	ANN has shown the highest predictive rate of accuracy (*i.e*. 89%) for diabetes diagnoses

Fig. 9 represents the predictive rate of accuracies achieved by different authors using soft computing techniques for diabetes diagnosis. Based upon data and feature set, the predictive accuracy lies between 74 to 100%.

Rate of Accuracy Achieved using Hybrid Algorithms: Fig. 10 represents the predictive rate of accuracies achieved using individual and hybrid GA by different authors for diabetes diagnosis. It is observed that the range of accuracy using hybrid GA lies between 74.78% - 98.8%.

Fig. 11 represents the predictive rate of accuracies achieved using individual and hybrid PSO by different authors for diabetes diagnosis. It is observed that the range of accuracy using hybrid PSO lies between 78.5% - 97.83%.

Fig. 12 represents the predictive rate of accuracies achieved using individual and hybrid ABC by different authors for diabetes diagnosis.

It is observed that the range of accuracy using hybrid ABC lies between 79.29% - 90%.

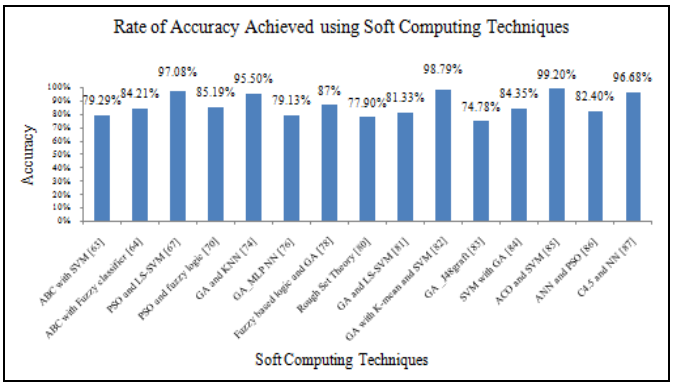

FIG. 9: RATE OF ACCURACY ACHIEVED USING SOFT COMPUTING TECHNIQUES

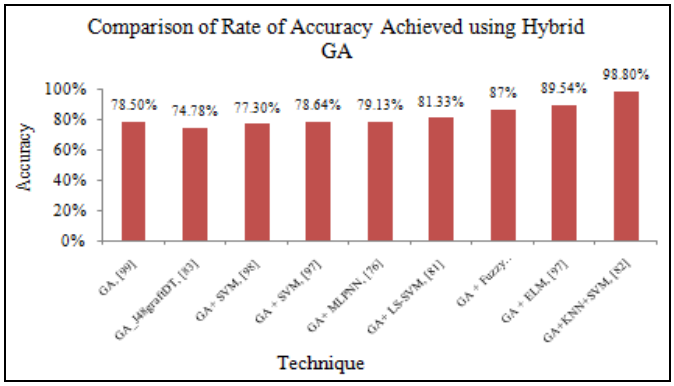

FIG. 10: COMPARISON OF RATE OF ACCURACY ACHIEVED USING HYBRID GA FOR DIABETES DIAGNOSIS

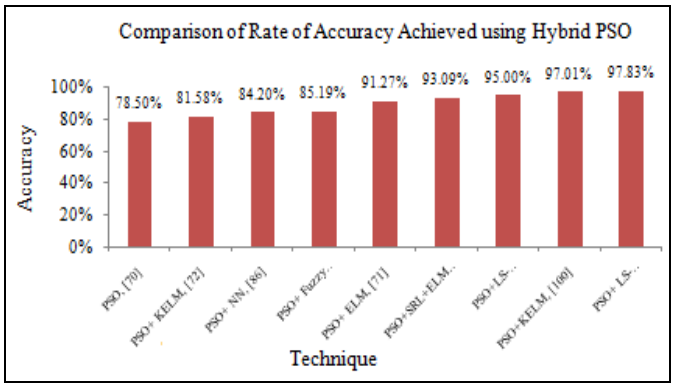

FIG. 11: COMPARISON OF RATE OF ACCURACY ACHIEVED USING HYBRID PSO FOR DIABETES DIAGNOSIS

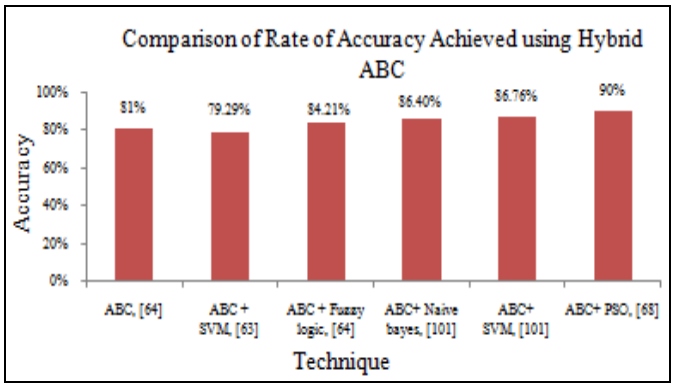

FIG. 12: COMPARISON OF RATE OF ACCURACY ACHIEVED USING HYBRID ABC FOR DIABETES DIAGNOSIS

It is evident that hybridization of traditional data mining and soft computing techniques is capable to produce more precise diagnostic results as compared to their individual use.

Rate of Accuracies Achieved by Analysing Image Data for Diabetes Diagnosis: Some of the key authors diagnose diabetes using iris or tongue image data and measured accuracy with different data mining and soft computing techniques. S. B More *et al*., (2015), diagnose type 2 diabetes in human being using irido-diagnosis. In the study, authors have used the iris image data, perform data pre-processing steps to normalize the data, and perform image segmentation to separate pupil and schlera data. SVM method was used to classify the image data. The rate of accuracy shown by the study was 90 to 92% for type 2 diabetes diagnosis [88]. A. Bansal *et al*., (2015), used SVM method to diagnose diabetes by using iridology.

The study used eye images features of 80 patients for diabetes diagnosis. The rate of accuracy given by the study was 87.50% [89]. P Samant *et al*., (2017), diagnose diabetes by analysing iris data of 200 patients using several data mining algorithms such as binary tree, random forest, SVM, neural networks. The study has shown the highest rate of accuracy (*i.e.* 89.66%) using random forest method [90]. J. Zhang *et al*., (2017), proposed a diagnostic model of diabetes using SVM and GA. Author diagnose diabetes on the basis of tongue images of patients by analysing the features like texture and colour tongue. Also, the rate of accuracy of proposed model was compared by other methods *viz.* KNN, naive bayes, BPNN, where the proposed model has shown the highest rate of accuracy (*i.e.* 79.72%) [91].

Fig. 13 shows the rate of accuracy achieved by analysing image data using different data mining

and soft computing techniques. It seems that depending upon selection of feature set used, the predictive rate of accuracy can be achieved between 67 to 92%.

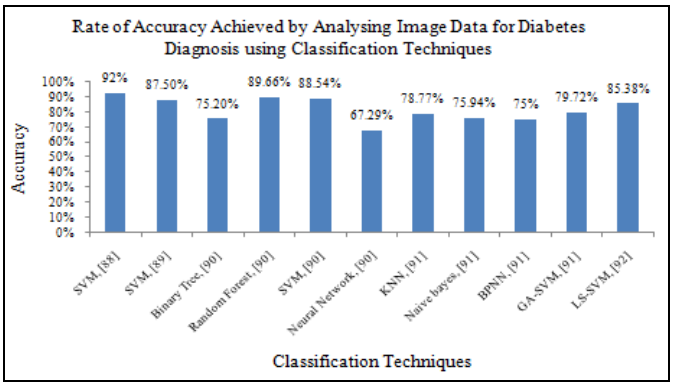

FIG. 13: RATE OF ACCURACY ACHIEVED BY ANALYSING IMAGE DATA USING DATA MINING AND SOFT COMPUTING TECHNIQUES

Effect of Pre-processing data: In healthcare informatics, data is collected from different sources and in heterogeneous formats.

Therefore, it may contain incomplete, noisy, redundant, or inconsistent values. Data pre-processing is used to remove these types of anomalies. Data cleaning, data reduction, normalization, deviation, detection are some of the important pre-processing techniques. It has been found that effective and innovative use of data pre-processing can improve accuracy rate of different data mining and soft computing techniques [93, 94].

Also, by analysing the article publications using Google scholar database, it is observed that usage of pre-processing methods for diabetes diagnosis increases regularly. **Fig. 14** shows the use of pre-processing methods for diabetes diagnosis in last ten years. Kandhasamy *et al.*, have analyzed the effect of pre-processing in diabetes diagnosis using J48, KNN, random forest and SVM. Authors found that rate of accuracies have improved by 10 - 25% [45]. **Fig. 15** shows the effect of pre-processing techniques on diabetes diagnosis procedure.

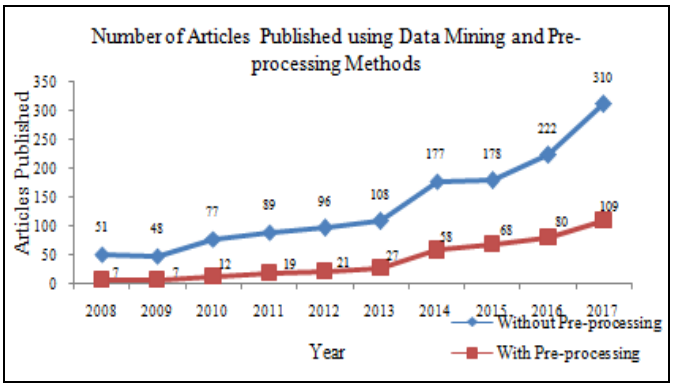

FIG. 14: NUMBER OF ARTICLES PUBLISHED USING DATA MINING AND PRE-PROCESSING METHODS

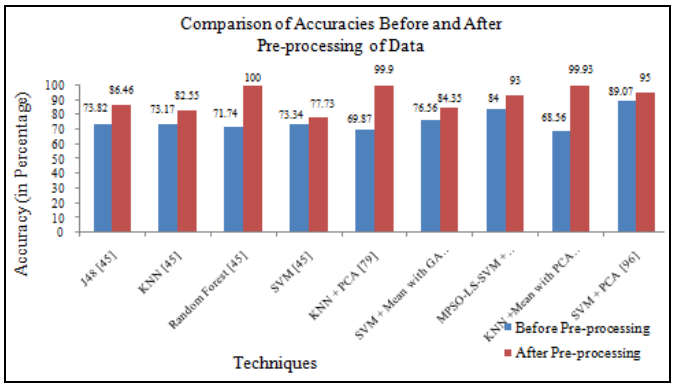

FIG. 15: EFFECT OF PRE-PROCESSING ON ACCURACY RATE

From **Fig. 15**, it is observed that the use of pre-processing techniques always improve the predictive rate of accuracy.

Proposed Framework: It is observed that considerable amount of effort has been already laid down for diabetes diagnosis. However, still smart and more advanced diabetes diagnostic solution is required which may offer the following facilities:

- ✓ Online data transmission between the different health related sources like hospitals, clinics, doctors, research centres *etc*.
- ✓ Save patients time and money which otherwise will be wasted in travelling and waiting.
- ✓ Privacy of patients data
- ✓ 24X7X367 connectivity to different health related sources
- ✓ Effective and user friendly diagnosis reports.
- ✓ Reduce readmission rate of diabetic patients
- ✓ Regular monitoring of diabetes patients.

As shown in **Fig. 16**, an innovative diabetes diagnosis framework based upon IoT, machine learning, emerging nature inspired computing techniques (Ant Lion Optimizer [102], Dolphin Echolocation [103], Firefly Algorithms [104], Grasshopper Optimization [105] *etc*) ontology and information theory is proposed. IoT will assist in collecting and communicating the diabetes patients between different sources like patients, hospital, clinics, embedded healthcare devices, laboratory

and research centres. The IoT security layer is provided to maintain the privacy of the patient's data. The use of ontologies and machine learning techniques will assist in getting better diagnosis results. The use of soft computing will help in expediting the diagnosis process. Information theory will try to avoid the slow convergence problem of the soft computing techniques. The system will also helps in reducing the readmission rate of diabetic patients.

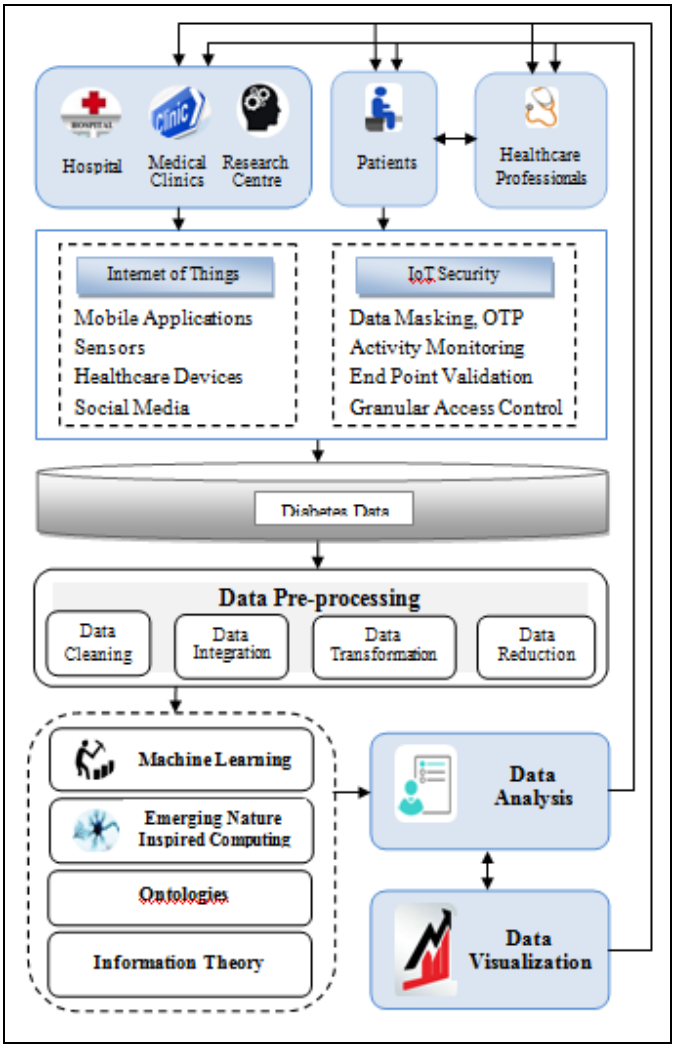

FIG. 16: PROPOSED FRAMEWORK FOR DIABETES DIAGNOSIS

CONCLUSION: Diabetes is a persistent human metabolic health disorder. Researchers have used different data mining (ID3, Random Forest, naive bayes, SVM, KNN, C4.5) and soft computing (GA, ACO, PSO, ABC) techniques to prospect diabetes disorder in human beings. The distinct data mining techniques are used to get the most promising consequences that represent high rate of success. In last ten year, the rate of usage of different data mining and soft computing techniques has been appreciably increased. Most of the researchers have used WEKA and MATLAB for early diagnosis of diabetes. In last 10 years, C4.5 and ID3 were the most and least preferred choices for mining diabetic patients. In soft computing techniques, the superlative research has been done using GA. However, very few researchers have implemented ABC for the same. The rate of accuracy of C4.5, SVM, KNN, random forest and hybrid approaches in diabetes diagnosis lies between 67 - 91%, 77 - 96%, 73% - 80%, 68 - 94% and 74 - 99.2% respectively. It seems that hybrid techniques are more effective and accurate as far as the rate of accuracy for diabetes diagnosis is concerned. The use of data pre-processing techniques can further improve the predictive rate of accuracy.

No doubt, significant work for diabetes diagnosis has been done. However, still smart and more advanced solution is required which may automatically collect and monitor different activities of diabetic patients. Therefore, an attention is required to develop a smart and hybrid diabetic diagnostic system using Machine Learning, Soft Computing, Internet of Things, Ontologies and Information Theory to awake and save human masses from extensive decisive spectrum of this deadliest human disorder. In addition, the system should be able to reduce the readmission rate of diabetic patients.

ACKNOWLEDGEMENT: We acknowledge all the researchers who worked on diabetes diagnosis using data mining and soft computing.

CONFLICT OF INTEREST: There is no conflict of interest.

REFERENCES:

1. Olivareza JP, Olivaresa FA, Medinaa AP, Carmonaa JD, Agundisa AR and Calderona AE: Bioimpedance phase angle analysis of foot skin in diabetic patients: Anexperimental case study. IRBM 2015; 36: 233-239.
2. Kaul K, Tarr JM, Ahmad SI, Kohner EM and Chibber R: Introduction to Diabetes Mellitus. Diabetes: An Old Disease, a New Insight. Springer-Verlag New York, Edition 1, 2013; 1-11.
3. Sharma M, Singh G and Singh R: Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques. IRBM 2017; 38: 305-324.
4. Kaveeshwar SA and Cornwall J: The current state of diabetes mellitus in India. Australas Med J 2014; 7: 45-48.

5. Wild S, Roglic G, Green A, Sicree R and King H: Prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care 2004; 27: 1047-1053.
6. Shaw JE and Simpson RW: Prevention of type 2 diabetes. Diabetes and Exercise: Human Press, 2009: 55-68.
7. IDF Diabetes Atlas - 8th Edition. Available from: http://www.diabetesatlas.org/across-the-globe.html [accessed on 31st December, 2017]
8. Diabetes.co.uk, the global diabetes community. Available from: https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwi9keWYsrPYAhWGNo8KHTzeDW4QFggqMAE&url=https%3A%2F%2Fwww.diabetes.co.uk%2Fdiabetes-prevalence.html&usg=AOvVaw00atd_9amFAf5SOi5qc2ht [accessed on 31st December, 2017]
9. Global Report on Diabetes. World Health Organization. 2016; 1-88. Available from: http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf [accessed on 28th December, 2017]
10. Anjali K: A Review on the Diagnosis of Diabetes Mellitus. International Journal of Digital Application and Contemporary Research 2015; 4(1): 1-7.
11. Doumbouya MB, Kamsu-Foguem B, Kenfack H and Foguem C: A framework for decision making on teleexpertise with traceability of the reasoning, IRBM 2015; 36: 40-51.
12. Available from: https://www.alz.co.uk/research/world-report-2016
13. Available from: https://www.salk.edu/news/salk-news/faq-on-type-4-diabetes/
14. Rubin DJ: Hospital Readmission of Patients with Diabetes, Hospital Management of Diabetes 2015; 15(17): 1-9.
15. Dungan KM: The Effect of Diabetes on Hospital Readmissions. J Diabetes Sci Technol 2012; 6(5): 1045-1052.
16. Silverstein MD, Qin H, Mercer SQ, Fong J and Haydar Z: Risk factors for 30-day hospital readmission in patients ≥65 years of age. Bayl Univ Med Cent 2008; 2(4): 363-372.
17. Marinov M, Mosa ASM, Yoo I and Boren SA: Data Mining Technologies for Diabetes: A Systematic Review. J diabetes Sci Technol 2011; 5: 1549-1556.
18. Verma P, Kaur I and Kaur J: Review of Diabetes Detection by Machine Learning and Data Mining. International Journal of Advance Research, Ideas and Innovations in Technology 2016; 2: 1-5.
19. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I and Chouvarda I: Machine Learning and Data Mining Methods in Diabetes Research, Comput. Struct. Biotechnol. J. 2017; 15: 104-116.
20. Deshmukh T and Fadewar HS: Data Mining Techniques for Diagnosis of Diabetes: A Review, International Journal of Emerging Research in Management and Technology 2017; 6(9): 212-214.
21. Okikiola FM, Mustapha AM, Akinade OA, Adeleye EO and Alonge CY: A Systematic Literature Review on Diabetes Diagnosis Management System. International Journal of Engineering And Computer Science 2017; 6(9): 22398-22315.
22. Pawar S and Smita S: An extensive survey on diagnosis of diabetes mellitus in healthcare. Advances in intelligent systems and computing 2017; 468: 97-104.
23. Jiawei H, Micheline K and Jian P: Data Mining: Concepts and Techniques. Elsevier, Third Edition 2011.
24. Ian W and Eibe F: Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Second Edition 2005.
25. Agarwal P and Mehta S: Nature-Inspired Algorithms: State-of-Art, Problems and Prospects. Int J Comput Appl 2014; 100(14): 14-21.
26. Mohamed NE and Morsi El-Bhrawy AS: Artificial Neural Networks in Data Mining. IOSR J Comput Eng 2016; 18(6): 55-59.
27. Kaur P and Sharma M: A Survey on using Nature Inspired Computing for Fatal Disease Diagnosis. International Journal of Information System Modeling and Design 2017; 8(2): 70-91.
28. Sharma M, Singh G and Singh R: Design and analysis of stochastic DSS query optimizer in a distributed database system. Egyptian informatics journal 2015; 17: 161-173.
29. Zhang Y, Agarwal P, Bhatnagar V, Balochian S and Yan J: Swarm Intelligence and Its Applications. The Scientific World Journal 2013; 1-3.
30. Rini DP, Shamsuddin SM and Yuhaniz SS: Particle Swarm Optimization: Technique, System and Challenges. Int J Comput Appl 2011; 14: 19-27.
31. Blum C: Ant colony optimization: Introduction and recent trends. Phys Life Rev 2005; 2: 353-373.
32. Repalli P: Prediction on Diabetes using data mining Approach. SCSUG 2011.
33. Abdullah AA, Ahamad MG and Siddiqui MK: Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences 2013; 25: 127-136.
34. Rajesh K and Sangeetha V: Application of Data Mining Methods and Techniques for Diabetes Diagnosis. Int J Innov Res Sci Eng Technol 2012; 2: 224-229.
35. Lakshmi KR and Kumar SP: Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability. Int J Sci Res Eng 2013; 4: 933-942.
36. Nagarajan S, Chandrasekaran RM and Ramasubramanian P: Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes. Int J Curr Res Acad Rev 2014; 2: 91-98.
37. Vijayan VV and Aswathy RK: Study of data mining algorithm for prediction and diagnosis of diabetes mellitus. Int J Comput Appl 2014; 95: 12-16.
38. Thangaraju P and Bharathi NG: Data Mining Approaches for Diabetes using Feature selection. International Journal of Computer Science and Information Technologies 2014; 5: 5939-5943.
39. Radha P and Srinivasan B: Predicting Diabetes by cosequencing the various Data Mining. Int J Innov Res Sci Eng Technol 2014; 1: 334-339.
40. Sankaranarayanan S and Pramananda PT: Diabetic prognosis through Data Mining- Methods and Techniques. In: International Conference on Intelligent Computing Applications 2014; 162-166.
41. Mounika M, Suganya SD, Vijayashanthi B and Anand KS: Predictive Analysis of Diabetic Treatment Using Classification Algorithm. International Journal of Computer Science and Information Technologies 2015; 6: 2502-2505.
42. Kumar NMS, Eswari T, Sampath P and Lavanya S: Predictive Methodology for Diabetic Data Analysis in Big Data. Procedia Comput Sci 2015; 50: 203-208.
43. Agicha K, Bhatia P, Badlani N, Menghrajani A and Tewari A: Survey on Predictive Analysis of Diabetes in Young and Old Patients. International Journal of Advanced Research in Computer Science and Software Engineering 2015; 5: 445-450.

44. Iyer A, Jeyalatha S and Sumbaly R: Diagnosis of Diabetes Using Classification Mining Techniques. In: International Journal of Data Mining and Knowledge Management Process 2015; 5: 1-14.
45. Kandhasamy JP and Balamurali S: Performance Analysis of Classifier Models to Predict Diabetes Mellitus, Procedia Comput Sci 2015; 47: 45-51.
46. Perveea S, Shahbaz M, Guergachi A and Keshavjee K: Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Comput Sci 2016; 82: 115-121.
47. Daghistani T and Alshammari R: Diagnosis of Diabetes by Applying Data Mining Classification Techniques, Int J Adv Comput Sci Appl 2016; 7: 329-332.
48. Teimouri M, Farzadfar F, Alamdari MS, Meshkini AH, Alamdari A, PA and Darzi ER.: Detecting Diseases in Medical Prescriptions Using Data Mining Tools and Combining Techniques. Iran J Pharm Res 2016; 15: 113-123.
49. Thirumal PC and Nagarajan N: Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - A Case Study. ARPN Journal of Engineering and Applied Sciences 2015; 10: 8-13.
50. Sa'di S, Maleki A, Hashemi R, Panbechi Z and Chalabi K: Comparison of Data Mining Algorithms in the Diagnosis of Type II Diabetes. International Journal on Computational Science & Applications 2015; 5: 1-12.
51. Kamble TP and Patil S: Diabetes Detection using Deep Learning approach. Int J Innov Res Sci Eng Technol 2016; 2: 342-349.
52. Hina S, Shaikh A and Sattar SA: Analyzing Diabetes Datasets using Data Mining. Journal of Basic and Applied Sciences 2017; 13: 466-471.
53. Selvakumar S, Kannan KS and Gothainachiyar S: Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques. Chemometr Intell Lab Syst 2017; 12(2): 183-188.
54. Kumari V and Chitra R: Classification of Diabetes Disease using SVM. International Journal of Engineering Research and Applications 2013; 2(3): 1797-1801.
55. Kumar VP and Lakshmi V: A data mining approach for prediction and treatment of diabetes disease. Int J Sci Invent Today 2014; 3(3): 73-9.
56. Vispute NJ, Kumar SD and Rajput A: An empirical comparison by data mining classification techniques for diabetes data set. Int J Comput Appl 2015; 131(2): 6-11.
57. Gregori D, Petrinco M, Bo S, Rosato R, Pagano E, Berchialla P and Merletti F: Using Data Mining Techniques in Monitoring Diabetes Care. The Simpler the Better. J Med Syst 2011; 35: 277-81.
58. Thirugnanam M, Kumar P and Srivatsan V: Improving Prediction Rate of Diabetes Using Fuzzy, Neural Network, and Case Based (FNC) Approach. Procedia Eng 2012; 38: 1709-1718.
59. Pangaribuan JJ and Suharjito: Diagnosis of Diabetes Mellitus Using Extreme Learning Machine. In: International Conference on Information Technology Systems and Innovation 2014; 33-38.
60. Ince T, Kiranyaz S and Pulkkinen J: Evaluation of global and local training techniques over feed-forward neural network architecture spaces for computer-aided medical diagnosis. Expert Syst Appl 2010; 37: 8450-8461.
61. Karaboga D and Ozturk C: Fuzzy clustering with artificial bee colony algorithm. Scientific Research and Essays 2010; 5: 1899-1902.
62. Sapna S, Tamilarasi A and Parvin MK: Implementation of Genetic algorithm in predicting diabetes. Int J Comp Sci 2012; 9: 234-240.
63. Uzer MS, Yilmaz N and Inan O: Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification. The Scientific World Journal 2013; 1-10.
64. Beloufa F and Chikh MA: Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer Method and Program in Biomedicine 2013; 112: 92-103.
65. Kovalan S, Mugilan S and Balamurugan S: Using Genetic Algorithm for efficient mining of Diabetic data. Int J Eng Res Technol 2014; 3: 613-616.
66. Lakshmi KV and Padmavathamma M: A Model to Predict Diabetes Based on Chromosomes Using Genetic Algorithm. International Journal of Innovative Research and Development 2014; 3: 367-372.
67. Soliman OS and AboElhamd E: Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine. International Journal of Computer Trends and Technology 2014; 8: 39-44.
68. Karthikeyan S and Christopher T: A Hybrid Clustering Approach using Artificial Bee Colony (ABC) and Particle Swarm Optimization. Int J Comput Appl 2014; 100: 1-6.
69. Sreedevi E and Padmavathamma M: A Threshold Genetic Algorithm for Diagnosis of Diabetes using Minkowski instance Method. Int J Innov Res Sci Eng Technol 2015; 4: 5596-5601.
70. Sahebi HR and Ebrahimi S: A Fuzzy Classifier Based on Modified Particle Swarm Optimization for Diabetes Disease Diagnosis. Geoinformatica 2015; 5: 11-17.
71. Subbulakshmi CV and Deepa SN: Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier. The Scientific World Journal 2015; 1-7.
72. Zhang L and Yuan J: Fault Diagnosis of Power transformers using Kernel based Extreme Learning Machine with Particle Swarm Optimization. Appl. Math. Inf. Sci. 2015; 9: 1003-1010.
73. Soliman OS and El-Hamd EA: A Chaotic Levy Flights Bat Algorithm for Diagnosing Diabetes Mellitus. Int J Comput Appl 2015; 111: 1-7.
74. Pavate A and Ansari N: Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques. In Fifth International Conference on Advances in Computing and Communications (IEEE) 2015; 371-375.
75. Sreedevi E and Padmavathamma M: Design and Development of Hybrid Genetic Classifier Model for Prediction of Diabetes. International Journal of Modern Trends in Engineering and Research 2016; 3: 260-265.
76. Choubey DK, Paul S. GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis. International Journal Intelligent Systems and Applications 2016; 1: 49-59.
77. Komi M, Li J, Zhai Y and Zhang X: In: 2nd International conference on Image, Vision and Computing. Application of Data Mining Methods in Diabetes Prediction 2017. IEEE Xplore.
78. Ephzibah EP: Cost Effective Approach on Feature Selection Using Genetic Algorithms and Fuzzy Logic For Diabetes Diagnosis. International Journal on Soft Computing 2011; 2: 1-10.

79. Divya, Chhabra R, Kaur S and Ghosh S: Diabetes detection using artificial neural networks and back-propagation algorithm. International Journal of Scientific and Technology Research 2013; 2(1): 9-11.
80. Varma KVSRP, Apparao A and Rao PVN: A Computational Intelligence Technique for Effective and Early Diabetes Detection using Rough Set Theory. Int J Comput Appl 2014; 95: 17-21.
81. Aishwarya S and Anto S: A Medical Decision Support System based on Genetic Algorithm and Least Square Support Vector Machine for Diabetes Disease Diagnosis. International Journal of Engineering Sciences & Research Technology 2014; 3: 4042-4046.
82. Santhanam T and Padmavathi MS: Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. Procedia Comput Sci. 2015; 47: 76-83.
83. Choubey DK and Paul S: GA_J48graft DT: A Hybrid Intelligent System for Diabetes Disease Diagnosis. International Journal of Bio-Science and Bio-Technology 2015; 7: 135-150.
84. Asgarnezhad R, Shekofteh M and Boroujeni FZ: Improving Diagnosis of Diabetes Mellitus using Combination of Preprocessing Techniques. J Theor Appl Inf Technol 2017; 95(13): 2889-2895.
85. Rashmi and Saini S: Hybrid Model Using Unsupervised Filtering Based on Ant Colony Optimization And Multiclass SVM by Considering Medical Data Set. International Research Journal of Engineering and Technology 2017; 4(6): 2565-2571.
86. Ateeq K and Ganapathy G: The novel hybrid Modified Particle Swarm Optimization - Neural Network (MPSO-NN) Algorithm for classifying the Diabetes. Comput Intell. 2017; 13(4): 595-614.
87. Peter S: An analytical study on early diagnosis and classification of diabetes mellitus. Bonfring Int J Data Min 2014; 4(2): 7-11.
88. More SB and Pergad ND: On a Methodology for Detecting Diabetic Presence from Iris Image Analysis. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2015; 4(6): 5234-5238.
89. Bansal A, Agarwal R and Sharma RK: Determining diabetes using iris recognition system. International Journal of Diabetes in Developing Countries 2015; 1-7.
90. Samant P and Agarwal R: Diagnosis of Diabetes using Computer Methods: Soft Computing Methods for Diabetes Detection Using Iris. International Journal of Biomedical and Biological Engineering 2017: 11(2): 63-68.
91. Zhang J: Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images. BioMed Research International 2017; 2017: 1-10.
92. Sakthivel K: A Support Vector Machines based Classifications of the Diabetes Mellitus from Human Tongue Quantitative Features. Asian Journal of Research in Social Sciences and Humanities 2016; 6(7): 584-592.
93. Razavi AR, Gill H, Åhlfeldt H and Shahsavar N: A Data Pre-processing Method to Increase Efficiency and Accuracy in Data Mining. AIME 2005; 434-443.
94. Vembandasamy K and Karthikeyan T: Novel outlier detection in diabetes classification using data mining techniques. Int J Appl Eng Res 2016; 11(2): 1400-3.
95. Jayalskshmi T and Santhakumaran: A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. International Conference on Data Storage and Data Engineering 2010.
96. Aishwarya R, Gayathri P and Jaisankar N: A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology 2013; 5(3): 2903-2908.
97. Aishwarya S and Anto S: A Medical Expert System based on Genetic Algorithm and Extreme Learning Machine for Diabetes Disease Diagnosis. International Journal of Science, Engineering and Technology Research 2014; 3(5): 1375-1380.
98. Kumar GR, Ramachandra GA and Nagamani K: An Effective Feature Selection System to Integrating SVM with genetic algorithm for large medical dataset. International Journal of Advanced Research in Computer Science and Software Engineering 2014; 4(2): 272-277.
99. Aslam MW and Nandi AK: Detection of Diabetes Using Genetic Programming. 18th European Signal Processing Conference, Denmark 2010.
100. Zhan L and Yuan J: Fault Diagnosis of Power Transformers using Kernel based Extreme Learning Machine with Particle Swarm Optimization. International Journal of applied Mathematics and Information Sciences 2015; 9(2): 1003-1010.
101. Subanya B and Rajalaxmi RR: Artificial Bee Colony based Feature Selection for Effective Cardiovascular Disease Diagnosis. International Journal of Scientific and Engineering Research 2014; 5(5): 606-612.
102. Mirjalili S: The Ant Lion Optimizer. Adv. In Engg. Soft. 2015; 83: 80-98.
103. Kaveh A and Farhoudi N: Dolphin Echolocation Optimization: Continuous search space. Advances in Computational Design 2016; 1(2): 175-194.
104. Yang XS: Firefly Algorithm, Lévy Flights and Global Optimization. Research and Development in Intelligent Systems XXVI. Springer 2010; 209-218.
105. Mirjalili SZ, Mirjalili S and Saremi S: Grasshopper optimization algorithm for multi-objective optimization problems. Applied Intelligence 2017; 1-16.

How to cite this article:
Kaur P and Sharma M: Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review. Int J Pharm Sci Res 2018; 9(7): 2700-19. doi: 10.13040/ IJPSR.0975-8232.9(7).2700-19.