Asian Journal of Information Technology 15 (17): 3217-3231, 2016
ISSN: 1682-3915

Computer Aided Diagnosis System for Clinical Decision Making: Experimentation Using Pima Indian Diabetes Dataset

[1]N. Leema, [1]H. Khanna Nehemiah, [2]A. Kannan and [1]J. Jabez Christopher
[1]Ramanujan Computing Centre
[2]Information Science and Technology, College of Engineering Guindy, Anna University,
600025 Chennai, India

Abstract: Diabetes is a major health problem, the society faces today. Diagnosis and treatment of diabetes will improve the quality of life of affected individuals. Clinical Decision Support System (CDSS) serves as an aid for junior clinicians to diagnose diabetes in the absence of an expert diabetologist. This research aims to develop a CDSS diagnose the presence or absence of Gestational Diabetes Mellitus (GDM). The framework used to develop the CDSS has three subsystems, namely preprocessing subsystem, training subsystem and classification subsystem. Noisy values are handled by the preprocessing subsystem. The training subsystem fuzzifies the preprocessed data and constructs the hidden nodes of the Radial Basis Function Neural Network (RBFNN) using the Gaussian Membership Function. The exact interpolation property of the RBFNN is used to extract the weights between the hidden layer and the output layer. During RBFNN training, each instance in the training set is considered as fuzzy rules. The extracted weights are used to prune the generated fuzzy rules. Finally, the pruned rules are stored in a knowledge base. The Fuzzy Inference System uses the rules from the knowledgebase to classify the samples in the testing set. The CDSS for Gestational Diabetes Mellitus attains an overall accuracy of 88.31% with 79.31% sensitivity and 93.75% specificity. Our CDSS yields comparable classification performance when compared to the works of other researchers in the past decade. The CDSS serves as a second source of opinion for junior clinicians for the diagnosis of GDM. The classification frameworks used in this CDSS can be adopted for other clinical datasets.

Key words: Clinical decision support system, pima Indian diabetes, gestational diabetes mellitus, radial basis function neural network, fuzzy inference system

INTRODUCTION

Diabetes is a major health problem and its incidences are increasing globally. Statistics regarding the impact of diabetes in the society can be obtained from International Diabetic Federation (http://www.idf.org). "Diabetes is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action or both" (American Diabetes Association, 2014). People affected with diabetes and maintain consistently high blood glucose levels have an increased risk leading to serious diseases which affects the heart, eyes, kidneys, nerves and teeth.

Commonly observed signs in diabetes patients include frequent urination, excessive thirst, increased hunger, weight loss, tiredness, lack of interest and concentration, a tingling sensation or numbness in the hands or feet, blurred vision, frequent infections, slow-healing wounds, vomiting and stomach pain. Sometimes a diabetic patient may not experience any of the above signs.

The common types of diabetes are type-1 (juvenile diabetes), type-2 (adult onset diabetes) and Gestational Diabetes Mellitus (GDM). There are also other miscellaneous types of diabetes (http://www.diabetes.org.uk/). An etiologic classification of diabetes mellitus is presented in (American Diabetes Association, 2014). The origin of type-1 diabetes is the autoimmune destruction of the insulin-producing beta cells in the pancreas and as a result of this the pancreas is unable to produce insulin. The lack of insulin leads to increased blood and urine glucose. A person can be affected with type-1 diabetes at any age but in the majority of the cases, it affects before the age of 40 especially in the childhood (http://www.diabetes.org.uk/). Type-2 diabetes occurs when insulin produced is insufficient and thus cannot be used by the human body to control the blood sugar levels. This is common after the age of 40, however, people in the age group between 25-40 has also been diagnosed with type-2 diabetes (http://www.idf.org). A person is affected with pre-diabetes when the blood glucose level is higher than the normal range but less than the threshold value

Corresponding Author: N. Leema, Ramanujan Computing Centre, College of Engineering Guindy, Anna University, 600025 Chennai, India

required to classify a patient as diabetic. A person affected with pre-diabetes has an increased risk of getting affected by type-2 diabetes (http://www.diabetes.org/).

In women, the period of pregnancy is called gestation period. GDM is a type of diabetes which affects pregnant women. Pregnant women are at a high risk of being affected by GDM based on multiple clinical factors namely, previous diagnosis of GDM, prediabetes, member of a high risk population, age ≥35 years, Body Mass Index (BMI) ≥30 kg/m^2, Polycystic Ovary Syndrome (PCOS), acanthosisnigricans, use of corticosteroid, history of macrosomic infant and current fetal macrosomiaorpolyhydramnios (http://www.diabetes.org/). If a pregnant woman is diagnosed with diabetes during the first trimester of pregnancy, it is said be type -2. GDM is a type of diabetes diagnosed during the second or third trimester of pregnancy (American Diabetes Association, 2014). A woman is affected by GDM when the body is unable to produce enough insulin to meet the extra needs of pregnancy. Women with high glucose level during pregnancy can lead to the fetus gain abnormal weight. This can lead to problems during delivery, trauma to the child and mother, a sudden drop in blood glucose for the child after birth and the child is at a higher risk of developing diabetes in the future (http://www.idf.org).

In this research, a Clinical Decision Support System (CDSS) that will aid a physician to diagnose the presence or absence of Gestational Diabetes Mellitus (GDM) has been developed. CDSS are computerized expert systems that are developed using the knowledge extracted from clinical data sets and expert judgement. MYCIN is an example of early CDSS developed in 1970's to assist the physician in identifying bacteria causing infection and to recommend antibiotics (Shortliffe *et al*., 1975). A diagnostic decision support system for adverse drug reaction using temporal reasoning (Nehemiah and Kannan, 2006), detection of Bronchietasis in chest computed tomography images (Elizabeth *et al*., 2009), allergic rhinitis based on intradermal skin tests has been developed in 2015 (Christopher *et al*., 2015) and a temporal mining framework for classifying un-evenly spaced clinical data for building effective clinical decision making (Jane *et al*., 2016).

Literature review: Related works carried out by researchers using the Pima Indian diabetes data sets from the UCI machine learning repository is discussed below. Polat and Gunes (2007a, b) have presented an expert system approach based on Principal Component Analysis (PCA) and Adaptive Neuro-Fuzzy Inference System (ANFIS) for diagnosing of diabetes disease. They have combined with two techniques, namely PCA and ANFIS. The PCA is used to reduce the dimension of the data set which means eight features were reduced to four features. This model was tested with PID data set obtained from the UCI machine learning repository and it achieved 89.47% accuracy.

Kahramanli and Allahverdi (2008) have designed a hybrid system for the diabetes and heart disease. This hybrid system is a combination of Fuzzy Neural Network (FNN) and ANN trained with the Back Propagation algorithm (BP). This model has four phases. They are standardization, fuzzification, fuzzified data was given to the input of ANN and defuzzification. The defuzzified crisp data were given to the input of the ANN and was trained by BP algorithm. They have tested this hybrid model with PID and Cleveland Heart Disease (CHD) data set obtained from the UCI machine learning repository. This model achieved 84.24 and 86.8% accuracy for PID and CHD dataset respectively.

Ghazavi and Liao (2008) have modeled fuzzy methods with selected features. They have used three methods to model the fuzzy classifier. They are namely, fuzzy K-Nearest Neighbor (KNN) algorithm, fuzzy clustering-based modeling and the adaptive network based fuzzy inference system. They have used twelve feature selection methods to extract the features for classification. This model was tested with Wisconsin Breast Cancer (WBC) and PID data set obtained from the UCI machine learning repository. The fuzzy KNN with mutual correlation based feature selection achieved 97.17% accuracy for WBC and ANFIS with selected features namely, plasma glucose, BMI and age achieved 77.65% and ANFIS with the features plasma glucose, DPF and age achieved 77.78% accuracy for PID dataset.

Temurtas *et al*. (2009) have presented a comparative study on diabetes disease diagnosis using neural networks. They used a Multilayer Neural Network (MLNN) structure trained using Levenberg-Marquard (LM) algorithm and a Probabilistic Neural Network (PNN) for diagnosing diabetes. They have used PID data set from the UCI machine learning repository. They have achieved a classification accuracy of 82.37% using MLNN with LM and 78.13% using PNN.

Lee and Wang (2011) have developed a fuzzy expert system for diabetes decision support application. They have used five layers fuzzy ontology for developing the fuzzy expert system to describe the knowledge with uncertainty. The five layer structure of the Fuzzy Diabetes Ontology (FDO) was developed and applied to the diabetes domain to model the diabetes knowledgebase and also developed the Semantic Decision Support System (SDSA). This was a combination of a knowledge construction mechanism, fuzzy ontology generating

mechanism and semantic fuzzy decision making mechanism. This fuzzy expert system was tested with PID data set obtained from the UCI machine learning repository. It achieved 77.3% accuracy for the PID data set.

Ganji and Abadeh (2011) have used Ant Colony Optimization (ACO) based classification system to extract a set of fuzzy rules for diagnosing of diabetes disease called FCS-ANTMINER. This model operates in two main stages namely, training and testing. During the training stage the ACO algorithm was applied to generate a set of fuzzy rules using training patterns. The combination of all these fuzzy rule sets form the fuzzy classification system. In the testing stage the performance of the classification system is tested using test patterns. This model was tested with PID data set obtained from the UCI machine learning repository. This achieves 84.24% accuracy for the PID data set.

Ephzibah (2011) have developed a model that combines Genetic Algorithm (GA) and fuzzy logic for feature selection and diagnosis of diabetes disease. They have used GA to solve the feature subset selection from a larger set of features. The selected feature subset is used in the fuzzy rule based classifier system. This model was tested with PID data set obtained from the machine learning repository. They have achieved 69 and 87% accuracy for fuzzy without GA and fuzzy with GA respectively.

Ghosh *et al*. (2014) have developed a neuro-fuzzy classification model for data mining. In the neuro-fuzzy classification technique, the inputs were fuzzified using generalized bell shaped membership function. This creates a fuzzification matrix for all input pattern associated with a degree of membership to different classes. The attribute class is based on the degree of membership value. This method was tested with ten benchmark data sets obtained from the UCI machine learning repository. They are WBC, KDD CUP 1999, statlog, satellite, Mammographic Mass, Wilt, Mushroom, PID, Iris, Spambase and Car evaluation. This method achieved 98.4, 98.7, 92.3, 84.4, 98.9, 99.8, 82.1, 96.7, 92.3 and 91.2% accuracies for the above data sets respectively.

Dennis and Muthukrishnan (2014) have presented an adaptive genetic fuzzy system for medical data classification. In this research the fuzzy systems' learning process was based on GA. The GA optimizes the rules and membership functions of the medical data. The fuzzy rules were generated from the data and the optimized rules are selected by the GA. The GA introduces a new operator called systematic addition. The fitness function was assigned as the frequency of occurrence of the rules in the training data. This model was tested with seven data sets namely, CHD, liver, iris, wine, PID, glass and mammogram obtained from the machine learning repository. The accuracy of this method was 86.6% for iris, 89% for wine, 89.80% for PID and 86.05% for glass data set.

Compared to the research works discussed in literature, the proposed system framework is different in the following ways: in this research the total number of fuzzified MFs are used as the number of neurons in the hidden layer. Second, to reduce the classification error, the exact interpolation property of RBFNN is used to extract the weights between the hidden and the output layer. These weights are summed and applied in the consequent part of the corresponding fuzzy rule. These weights are used to prune the generated fuzzy rules. These pruned rules are stored in the Knowledgebase. These rules are used by the FIS for classifying test instances.

MATERIALS AND METHODS

In this study an outline of the materials and methods namely, Radial Basis Function Neural Network (RBFNN) and Fuzzy Inference System (FIS) are discussed. These materials and methods play a major role in the proposed framework.

Data set description: The proposed framework has been experimented with Pima Indian diabetes (PID) data set from the UCI machine learning repository (Blake and Merz, 1998). The PID is the result of a research survey carried out in the National Institute of Diabetes and Digestive and Kidney Diseases, United States. This group is the one of the highest known rates of diabetes worldwide. This dataset all patients were females having age >21 years old of Pima Indian Heritage (Table 1).

The PID data set has 768 samples with eight features namely, number of times pregnant, plasma glucose tolerance test (mg/dL), diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), diabetes

Table 1: Presents the description of PID dataset

Pima Indian diabetes dataset				
Atribute	Center (μ)	SD (σ)	Min. value	Max. value
Preg	3.800	3.400	0	17
Plas	120.9	32.00	56	198
Pres	69.10	19.40	24	110
Skin	20.50	16.00	7	63
Insu	79.80	115.2	14	846
Mass	32.00	7.900	18.2	67.1
Pedi	0.500	0.300	0.085	2.42
Age	33.20	11.80	21	81

pedegree function and age (years). A class label is associated with each sample to indicate whether the individual is affected with Gestational diabetes or not. These samples were collected during the first trimester of pregnancy. Among 768 samples collected 268 samples (34.9%) have been diagnosed with Gestational diabetes and 500 samples (65.1%) without Gestational diabetes.

There are no missing values in the data set, however, there are 652 number of zeros in the data set. The features namely, number of times pregnancy, plasma glucose concentration, diastolic pressure, triceps skin fold thickness, 2hr serum insulin and body mass index has the value 0, for 111, 5, 35, 227, 374 and 11 instances. A total of 432 instances have one or more features with the value 0 associated with it. Gestational diabetes occurs during the second and third trimester of pregnancy, hence number of times pregnancy value must be atleast once. A total of 432 instances have one or more features with the value 0 associated with it. The value 0 corresponding to these features with respect to a sample is considered as a noisy value in the dataset.

Fuzzy inference system: In the FIS the crisp input x_i is used to determine the degree of input membership functions. A fuzzy set is characterized by its membership function ($\mu A_j(x_i)$). Each input node x_i is connected to the corresponding membership function $\mu A_{j(x_i)}=\{\mu A_1(x_i),\mu A_2(x_i)...\mu A_s(x_i)\},\forall j=1,2....s$ where s is the total number of membership functions for each input x_i. The membership value is between 0 and 1. This maps membership grade of each element of X to the membership function. The linguistic label is given for each membership function. The degree of membership function and the linguistic label is shown in Table 2.

Table 2: Level chart, fuzzy set and membership value for every attribute in PID data set

Attribute value	Level chart	Fuzzy set	σ	Mambership
Pima Indian diabetes data set				
Preg [1 17]	<3	Low		1.5
	3-6	Medium	3.4	4.9
	>6	High		11.5
Plas (glu) [56 198]	<70	Low		60
	70-155	Desirable	32.0	112.5
	>155	High risk		176.5
Diastplic press [24 110]	<80	Normal		50
	80-90	Prehypertense	19.4	85
	>90	Hypertension		100
Skin [7 63]	<12	Thin		9.5
	12-30	Medium	16.0	21
	>30	Thick		47.5
Insu [14 846]	<140	Low		77
	140-275	Medium	115.2	207.5
	>275	High		562.5
Mass [18.2 67.1]	<18.5	Low		18.35
	18.5-30	Normal	7.9	24.25
	>30	Overweight		48.55
Pedi [0.085 2.42]	<0.35	Low		0.175
	0.35-0.8	Medium		0.575
	>0.8	High	0.3	0.81
Age [21 81]	<35	Young		27.5
	35-45	Middle		40
	45-60	Old		52.5
	>60	Very old	11.8	75

Fuzzy antecedent part: The fuzzy AND operator is used to get the antecedent part of the fuzzy rule and is computed as follows

$$R = \Pi_{i=1}^{D} \mu A_s(x_i) \quad (1)$$

Where $\mu A_s(x_i)$ is the membership function of the corresponding input x_i.

Fuzzy consequents part: The output of the system y^p is composed of the consequent part of the fuzzy rule. The output is computed using the extracted weights from the RBFNN. These extracted weights are added and applied to the consequent part of the corresponding fuzzy rule.

RBFNN: RBFNN is a quadratic neural network with three functional distinct layers namely, the input layer, the hidden layer and the output layer. The hidden layer maps a nonlinear transformation of the input unit space to a higher-dimensional hidden-unit space and uses the Gaussian activation function. The output of the hidden node (Broomhead and Lowe, 1988) is computed using Eq. 2:

$$\varphi_j = \exp\left(\frac{\left\| x_i - \mu_j \right\|^2}{2\sigma_j^2}\right),\ j = 1, 2, ..., h \quad (2)$$

Where:

φ_j = The jth output of the hidden layer, x_i is the input vector of dimension D

μ_j = The jth center vector with the same dimension as the input vector

σ_j = The jth width which is scalar and h is the number of nodes in the hidden layer

The output layer maps a linear transformation from the higher dimensional hidden-unit space to the lower dimensional output unit space. Its output is calculated using the mathematical model presented as:

$$y^p = \sum_{j=1}^{h} w_{kj}\varphi_j(x^p) \quad (3)$$

Where:

y^p = The output of the RBFNN

x^p = The input pattern

w_{kj} = A weight matrix from the hidden layer to the output layer

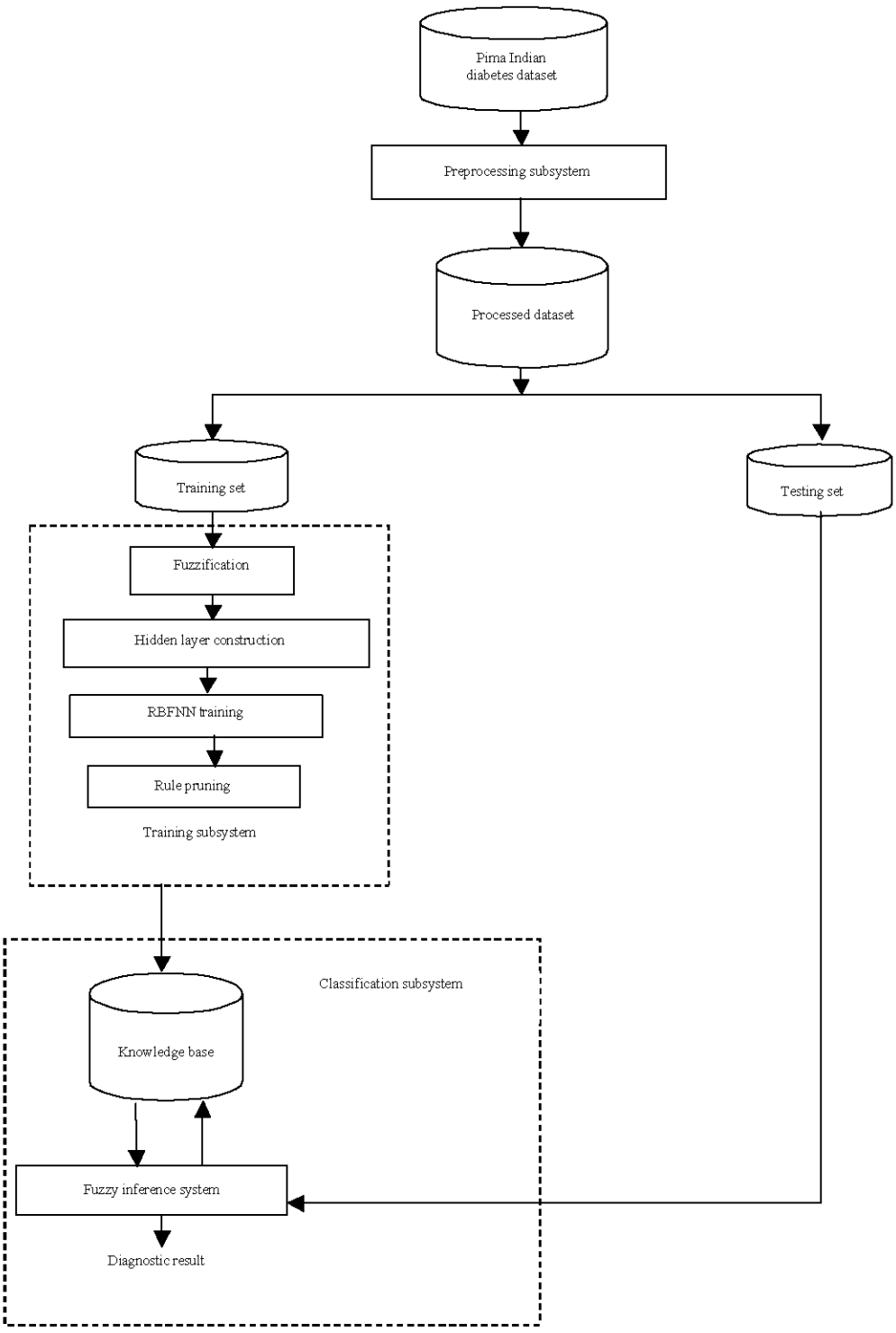

Fig. 1: CDSS system framework

System framework: The framework of the CDSS used in this research is illustrated in Fig. 1. The framework has three subsystems namely preprocessing subsystem, training subsystem and classification subsystem.

Preprocessing subsystem: The clinical data sets available in the UCI machine learning repository have both noisy and missing values. The data set used in this research is Pima Indian Diabetes. There are noisy values in the PID data set. This has the value 0 associated with it.

Smooth noisy data: Among 768 samples in the PID data set 432 instances have one or more features with the value 0 associated with it. If two or more features corresponding to an instance have the value 0 the sample is rejected. The number of instances with two or more features having the value 0 associated with it is 256. The number of instances in the PID data set is reduced from 768-512. Among the 512 instances 176 instances have the value 0 for one feature. The value zero is replaced by frequently occurring values of an attribute belongs to that class. From the 512 instances, 343 instances indicate the absence of diabetes (class 0) and 169 instances indicate the presence of diabetes (class 1).

Training subsystem: In the training subsystem, the preprocessed training inputs are fuzzified using Gaussian MF which is based on two parameters namely, center (μ) and width (σ). The fuzzified input features are characterized by its fuzzy set. The number of MFs in the fuzzy set is used to construct the hidden nodes in the hidden layer of RBFNN. The fuzzy set and its corresponding membership value for each feature in the PID dataset is presented in Table 2.

The training subsystem uses RBFNN which consists of three layers namely, input layer, hidden layer and the output layer. The input layer corresponds to each feature. The RBFNN training is a two phase training method; determination of number of neurons in the hidden layer and the weights between the hidden and the output layer. The number of neurons in the hidden layer is determined from the number of MFs in the fuzzy set. Hence, the hidden node describes the fuzzy partition of the input space which consists a fuzzy subset corresponding to each input feature. Hidden layer implements the antecedent part of the FIS. The center (μ_j) and width (σ_j) value for each hidden node is same as the MF corresponding to each input feature. The hidden unit activations $\varphi_j(x,\mu_j, \sigma_j)$ are fixed. There are totally 25 numbers of hidden nodes in the hidden layer. The second phase training determines the weights between the hidden and the output layer using the exact interpolation property of the RBFNN. This property uses pseudo-inverse technique to extract the weights between the hidden layer and the output layer. The output layer with output nodes corresponds to the class label. These extracted weights are summed and applied in the consequent part of the corresponding rules. These weights are used to prune the fuzzy rules from the training set. This subsystem prunes 98 rules from training dataset and stored in the knowledgebase. A set of sample rules are:

- If Preg Low and Glu Desirable and BP Normal and Skin Thickness Medium and Insulin Low and BMI Obese and DPF Low and Age Young then 0.03 Gestational Diabetes
- If Preg Low and Glu Desirable and BP Normal and Skin Thickness Medium and Insulin Low and BMI Overweight and DPF Low and Age Young then 0.05 Gestational Diabetes
- If Preg Low and Glu High Risk and BP Normal and Skin Thickness Thick and Insulin Low and BMI Obese and DPF Low and Age Young then 0.09 Gestational Diabetes
- If Preg Low and Glu Desirable and BP Normal and Skin Thickness Medium and Insulin Medium and BMI Normal and DPF Medium and Age Young then 0.03 Gestational Diabetes
- If Preg Medium and Glu High Risk and BP Normal and Skin Thickness Thick and Insulin Medium and BMI Obese and DPF High and Age Young then 0.04 Gestational Diabetes
- If Preg Medium and Glu High Risk and BP Normal and Skin Thickness Medium and Insulin Low and BMI Obese and DPF Low and Age Young then 0.04 Gestational Diabetes
- If Preg Low and Glu High Risk and BP Normal and Skin Thickness Medium and Insulin Low and BMI Obese and DPF Medium and Age Young then 0.03 Gestational Diabetes
- If Preg Low and Glu High Risk and BP Normal and Skin Thickness Medium and Insulin Low and BMI Overweight and DPF Medium and Age Young then 0.07 Gestational Diabetes
- If Preg Medium and Glu Desirable and BP Normal and Skin Thickness Medium and Insulin Low and BMI Obese and DPF Low and Age Young then 0.03 Gestational Diabetes

Classification subsystem: The classification subsystem uses the knowledgebase for classifying the gestational diabetes mellitus. This subsystem builds the classification model for developing the CDSS. The algorithm of the developed CDSS model is shown here. The CDSS used a FIS for classification of samples from the test dataset. The inference system interacts with the knowledgebase and provides the diagnostic result.

Steps in training the CDSS: The L number of training samples with i number of input attributes, $x^p = \{x_i^p, i = 1, 2, ..., D\}$ with D-dimensional input patterns has to be

mapped onto the corresponding target output y^p. The goal is to find the function f such that:

$$f(x^p) = y^p \quad \forall p = 1, 2, ...L \tag{4}$$

Step 1: The preprocessed inputs $x^p = \{x_i^p, i = 1, 2, ..., D\}$ are fuzzified by applying Gaussian MF $\mu A_j(x_i)$ and is computed using Eq. 2.

Step 2: The MF matrix of order D×S is generated which consist of the degree of memberships of D different patterns to S different classes. Each element in this matrix is a MF of the form $\mu A_j(x_i)$, those MFs forms the hidden nodes in the hidden layer. where x_i is the ith pattern value of input pattern vector x with i = 1, 2, ...D and j = 1, 2, ...S for fuzzification.

Step 3: Initialize the RBFNN parameters like learning rate (η) and the number of hidden nodes (h). Each instance in the training set is considered as a rule. The instances are used to train the RBFNN.

Step 4: The output weights are computed using the exact interpolation property of the RBFNN. The overall output is a simple linear combination of the hidden unit activations and is computed using Eq. 3 and this can be written as:

$$y^p = \varphi_j W_{jk} \tag{5}$$

Where y^p is the target output.

Step 5: The error of the network is computed using the following mathematical Eq. 6:

$$E = \frac{1}{L}\Sigma_{p=1}^{L}(y^p - O^p)^2 \quad \forall p = 1, 2, 3....L \tag{6}$$

Step 6: when the MSE (E) is minimum or zero, the exact interpolated weights are computed analytically as follows:

$$W = (\varphi^T\varphi)^{-1}\varphi^T y^p, \ \varphi' = (\varphi^T\varphi)^{-1}\varphi^T \tag{7}$$

$$W = \varphi' y^p \tag{8}$$

where, φ' is the pseudo-inverse of φ. The weights between the hidden and the output layer is computed using fast linear matrix inversion technique.

Step 7: These extracted weights are summed and applied to the consequent part of the corresponding fuzzy rule.

Step 8: The rules are pruned whose weights are lesser than τ, where τ is a user defined threshold $0 \leq \tau \leq 1$. In our experiment based on the classification accuracy of CDSS, we have set $\tau = 0.02$.

Step 9: The list of 98 pruned fuzzy rules is stored in the knowledgebase and used in the CDSS system for diagnosing of gestational diabetes disease.

Output: $R_i = \{R_1, R_2....R_n\}$, n = 1, 2, 3, ..., 98 where R_i is the set of rules.

RESULTS AND DISCUSSION

The performance of the CDSS is evaluated using PID data set obtained from the UCI machine learning repository. The classification results of the proposed CDSS performance are compared with the existing methods in the literature for PID data set. The results are analyzed based on classification accuracy and diagnostic test performance.

Observations and findings: This section discusses the implementation and comparative analysis of CDSS for gestational diabetes using PID dataset to observe the relationships between the input features and the corresponding output. In Fig. 2 the Gaussian MF for the PID data set is shown. The PID data set MF for:

- Number of times pregnant
- Plasma glucose tolerance test
- Diastolic blood pressure
- Triceps skin fold thickness
- Serum insulin
- Body mass index
- Diabetes pedigree function
- Age and
- Class label using different number of membership functions to describe the corresponding input and output features.

The MF value ranges from 0-1. The width of the MFs is based on the standard deviation (σ) of the corresponding input and output feature. The developed

Table 3: Diagnostic test evaluation of PID data set

		95% confidance interval	
Performance measures	PID	From	To
Sensitivity (%)	79.31	60.28	92.02
Specificity (%)	93.75	82.80	98.69
PLR	12.69	4.18	38.56
NLR	0.22	0.11	0.45
Disease prevalence (%)	37.66	26.87	49.40
PPV (%)	88.46	69.85	97.55
NPV (%)	88.24	76.13	95.56
MR	11.68	3.24	35.65
Accuracy (%)	88.31	68.45	96.61

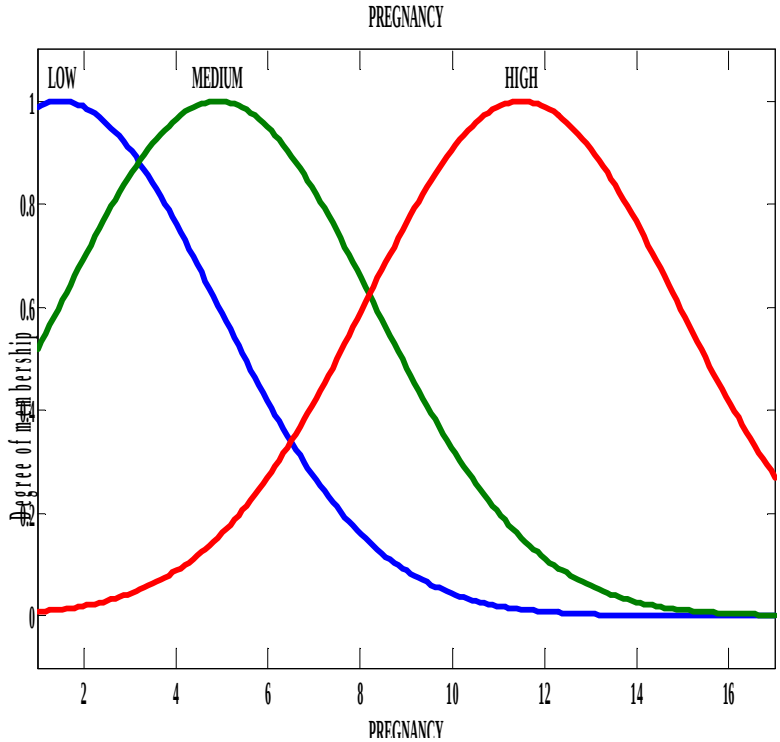

Fig. 2: Gaussian MF for pregnancy

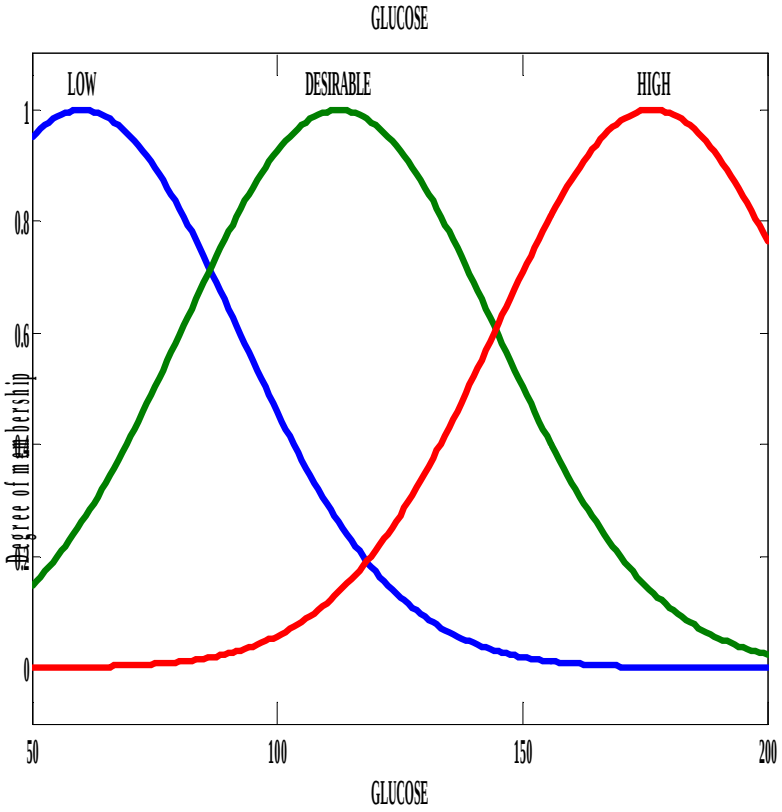

Fig. 3: Gaussian MF for glucose level

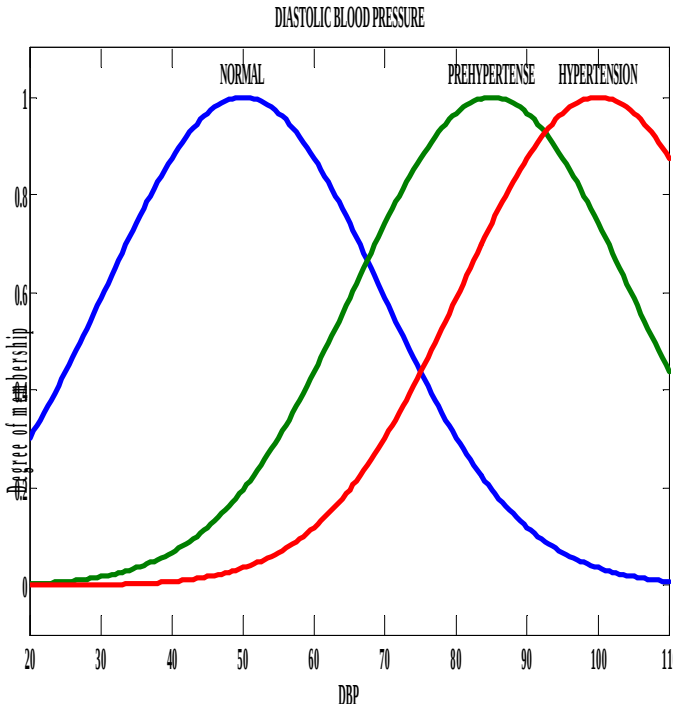

Fig. 4: Gaussian MF for blood pressure

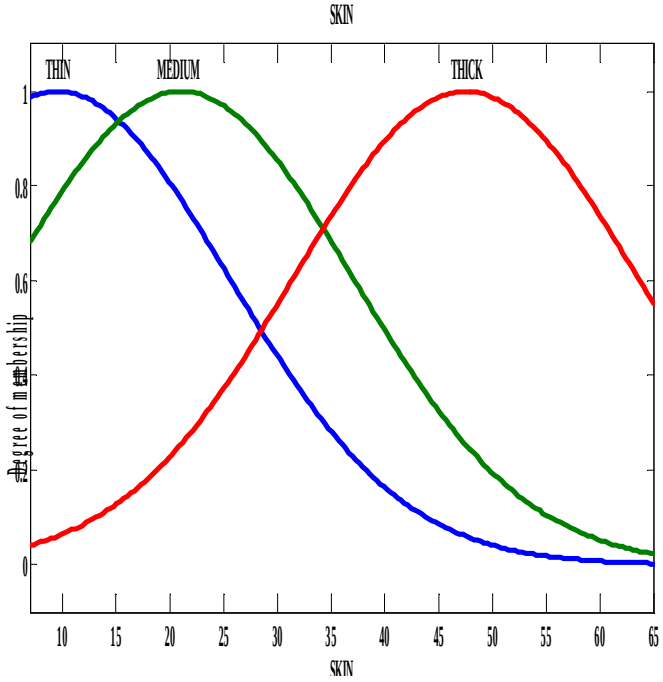

Fig. 5: Gaussian MF for skin fold thickness

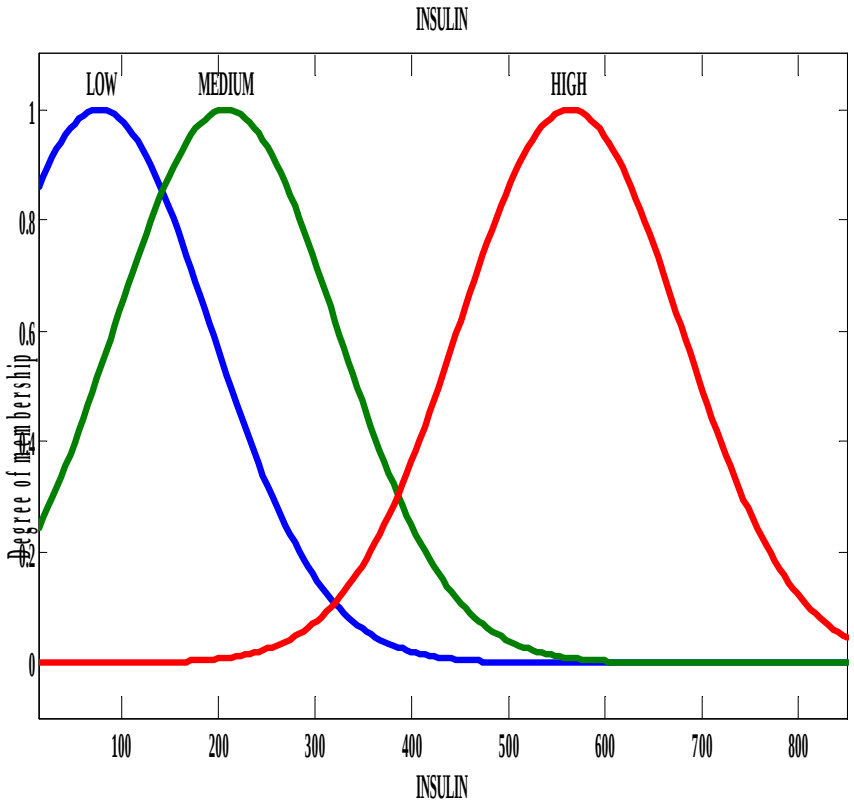

Fig. 6: Gaussian MF for serum insulin

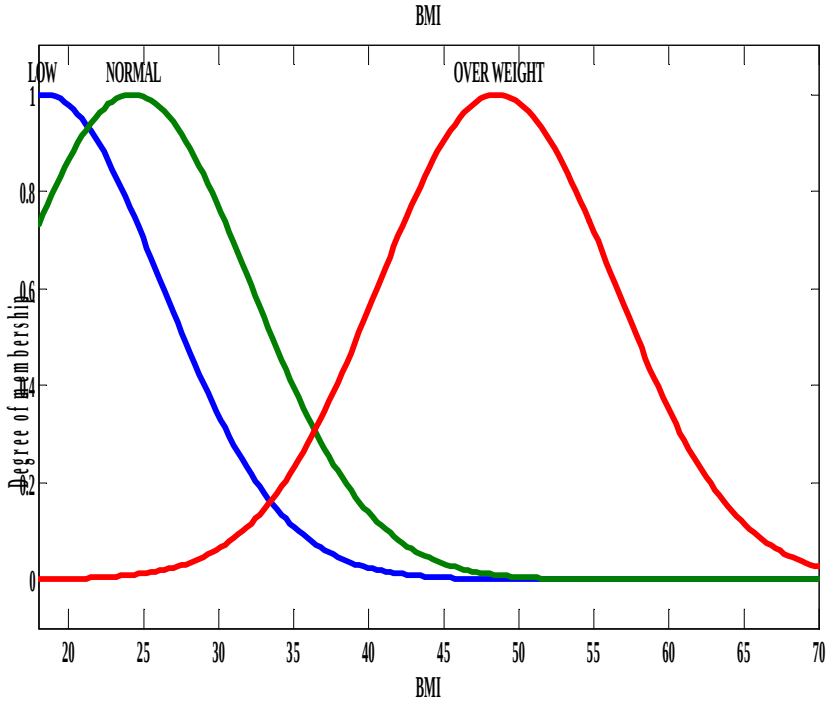

Fig. 7: Gaussian MF for BMI

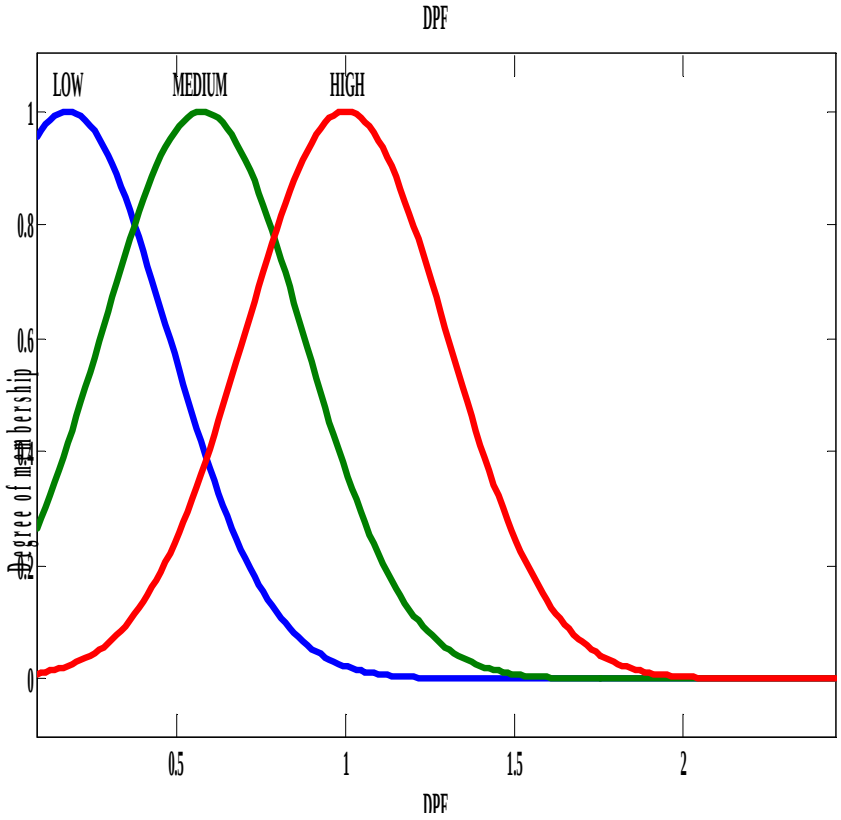

Fig. 8: Gaussian MF for DPF

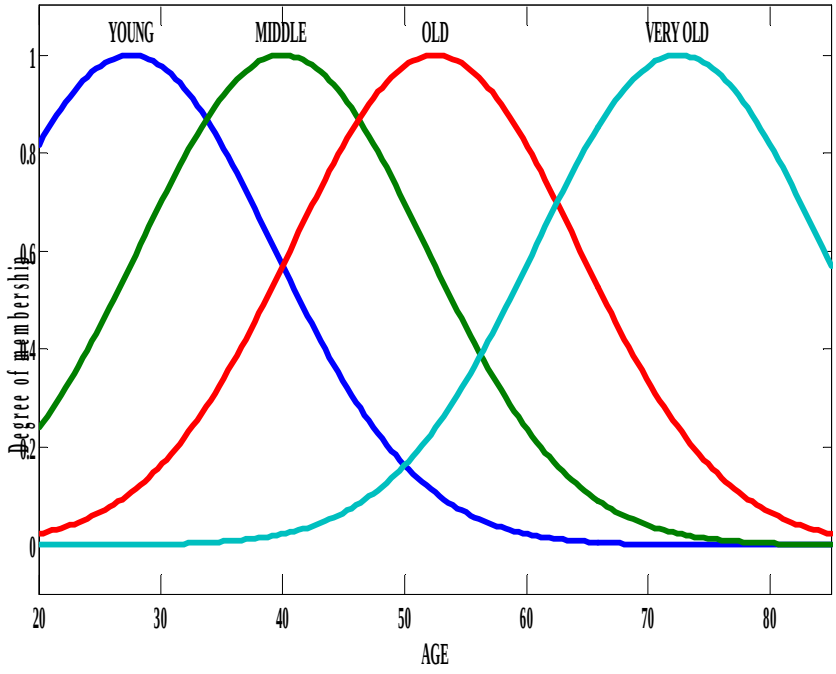

Fig. 9: Gaussian MF for AGE

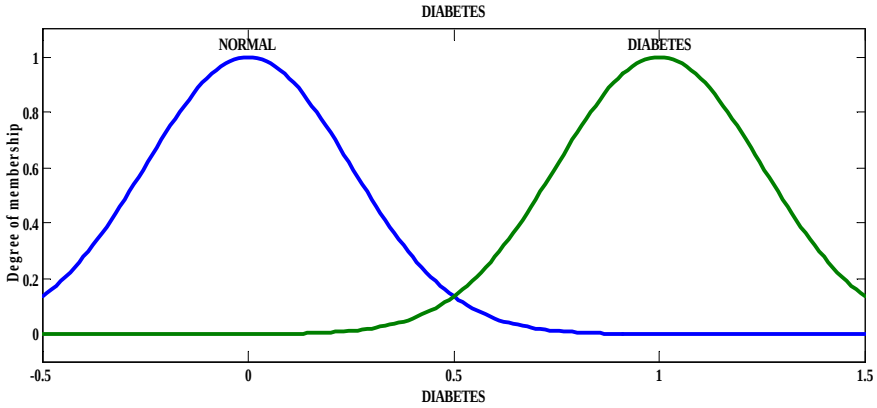

Fig. 10: Gaussian MF for diabetes level

Table 4: Copmarison accuracy of the proposed and some methods in literature for PID dataset

Author	Methods/Refrence	Accuracy/PID
Beloufa and Chikh (2013)	MABC (all features)	82.68
	MABC (features selection)	84.21
	ABC (all features)	79.61
	ABC (features selection)	81.40
Ganji and Abadeh (2011)	FCS-ANTMINER	84.24
Guzaitis *et al.* (2009)	GA-fuzzy classifier (10×CV)	71.49
Polat and Gunes (2007a b)	GDA-LSSVM (10×CV)	79.16
	LS-SVM	78.21
	C4.5	67.0±2.9
	MLP+BP	75.8±6.2
	Smart	76.8
	Linear discriment analysis	77.5
	QDA	59.5
	SNBa	75.4
	DIPOL92	77.4
	Semi-naive bayes	76.7±0.8
	OCN2	65.1±1.1
	MMLa	75.5±6.3
	KNN	71.7±6.6
	MML	75.5±6.3
	IB3	71.7±5.0
Palat	Fuzzy-AIRS (10×CV)	84.42
Tang and Tseng (2009)	BGA-fuzzy-KNN (5×CV)	81.6
	RGA-fuzzy-KNN (5×CV)	82
Palat	Neuro-fuzzy inference system (10×CV)	89.47
Temurtas *et al.* (2009)	ML-NN with LM (10×CV)	79.62
	ML-NN with LM	82.37
	PNN (10×CV)	78.05
Ghazavi and Liao (2008)	Fuzzy modeling (10×CV)	77.65
Lekkas and Mikhailov (2010)	Class, buffer capacity = 30	79.37
Goncalves *et al.* (2006)	Inverted Hierachical Neuro-Fuzzy system	78.60
Luukka and Leppalampi (2006)	Fuzzy similarity classifier	75.29
Sahan *et al.* (2005)	AWAIS	75.87
Elgawi and Hasegawa (2007)	Incremental Random Forests	76.80
Park and Choi (2008)	Incremental PCA	68.10
Cetisli (2010)	ANFC (9 fuzzy rules)	78.12
	ANFC-LH (fuzzy rules)	79.68
	ANFC (15 fuzzy rules)	79.16
	ANFC-HV (15 fuzzy rules)	80.72
Kahramanli and Allahverdi (2008)	Logdisc	77.7

Table 4: Continue

Author	Methods/Refrence	Accuracy/PID
	CART	72.8
	BP	75.2
	KNN, k = 23, Manh raw, W	76.7±4.0
	KNN, k = 1:25, Manh, raw	76.6±3.4
	ASR	74.3
	SSV DT	73.7±4.7
	Fisher Dicrimination analysis	76.5
	LFC	75.8
	ID3	71.7±6.6
	DB-CART	74.4
	Kohonen	72.7
	Hybrid system	84.2
Witten and Frank (2005)	Regression Cofficients	72.39
	Bayes	72.2±6.9
	C4.5	76.0±0.9
	NNGE	73.56
	RBF	75
	Naive Bayes	75.30
	SVM	76.0±0.9
	LWL	71.22
	Bayesian logistis regression	072.39
	ESOM	78.4
	IncNet	77.6
Mohamadi *et al.* (2008)	SA	75.71±4.41
Jaganathan *et al.* (2007)	ANTMINER	70.99
	ANT MINER with Imp. Qick Reduct	76.58
Quteishat *et al.* (2010)	Fuzzy-GA	89.74
Pulkkinen and Koivisto (2008)	Fuzzy classifier	78.22
Dennis and Muthukrishnan (2014)	AGFS	89.80
Chang	VISIT	77
Duch	C-MLP2LN	77.7
	C-MLP2LN	75.0
Ghosh *et al.*(2014)	NFS	82.1
	RBFNN	79.7
	ANFIS	80.5
Proposed	RBFNN-FIS	88.31

gestational CDSS consists of eight input neurons in the input layer, twenty five hidden neurons in the hidden layer and two output neurons in the output layer. Performance evaluation: the classification accuracy of the proposed model is evaluated by 10 fold cross validation

Table 5: The inference from the finding

								Diabetes	
Preg	Glu	DBP	Skin	Insulin	BMI	DPF	Age	From	To
12.14↑	158.6↑	-	-	-	-	-	-	0.673	0.730
12.14↑	-	97.14↑	-	-	-	-	-	0.662	0.724
13.36↑	-	-	7↑	-	-	-	-	0.652	0.727
7.28↑	-	-	-	432↑	-	-	-	0.650	0.780
9.71↑	-	-	-	-	25.43↑	-	-	0.648	0.722
4.85↑	-	-	-	-	-	1.099↓	-	0.693	0.763
1.214↑	-	-	-	-	-	-	25.14	0.691	0.769
-	158.6↑	97.14↑	-	-	-	-	-	0.604	0.617
-	-	97.14↑	15.29↑	-	-	-	-	0.583	0.636
-	-	77.86↑	-	133.4↓	-	-	-	0.563	0.694
-	-	90.71↑	-	-	32.86↑	-	-	0.604	0.648
-	-	97.14↑	-	-	-	1.267↓	-	0.583	0.727
-	-	90↑	-	-	-	-	26↑	0.677	0.720
-	158.6↑	-	11.14↑	-	-	-	-	0.583	0.636
-	153.8↑	-	-	133.4↓	-	-	-	0.5	0.671
-	158.6↑	-	-	-	32.86↑	-	-	0.583	0.642
-	158.6↑	-	-	-	-	1.267	-	0.583	0.727
-	150.3↑	-	-	-	-	-	23.43	0.720	0.769
-	-	-	15.29↑	312.6	-	-	-	0.521	0.994
-	-	-	7↑	-	47.71↑	-	-	0.583	0.648
-	-	-	7↑	-	-	1.099	-	0.583	0.746
-	-	-	7↑	-	-	-	23.79	0.677	0.769
-	-	-	-	133.4	40.29↑	-	-	0.5	0.597
-	-	-	-	133.4	-	1.436	-	0.538	0.773
-	-	-	-	133.4	-	-	27.5	0.563	0.777
-	-	-	-	-	32.68↑	1.099	-	0.659	0.728
-	-	-	-	-	29.14↑	-	23.29	0.648	0.769
-	-	-	-	-	-	0.760	24.21	0.727	0.788

of training and testing samples. The True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values are obtained from the testing set. The TP is the diseased sample correctly diagnosed as a disease. The FP is the normal sample incorrectly identified as a disease. The TN is a normal sample correctly identified as normal. The FN is a diseased sample incorrectly identified as normal.

The diagnostic test evaluation has been performed using (http://www.medcalc.org/calc/diaphpgnostic_test.php). The 95% Confidence Interval (CI) was also calculated. CI is a measure of the reliability of an estimate. The measures are computed as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (9)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (10)$$

$$\text{The Positive Likelihood Ratio(PLR)} = \frac{\text{TP rate}}{\text{FP rate}} = \frac{\text{Sensitivity}}{\text{1-Specificity}} \quad (11)$$

$$\text{The Negative Likelihood Ratio(NLR)} = \frac{\text{FN rate}}{\text{TN rate}} = \frac{\text{1-Sensitivity}}{\text{Specificity}} \quad (12)$$

$$\text{The Positive Predictive Ratio (PPV)} = \frac{TP}{TP + FP} \quad (13)$$

$$\text{The Negative Predictive Ratio (NPV)} = \frac{TN}{FN + TN} \quad (14)$$

$$\text{Misclassification Rate (MR)} = \frac{FP + FN}{(TP + FP + TN + FN)} \quad (15)$$

$$\text{Accuracy} = \frac{TP + TN}{(TP + FP + TN + FN)} \quad (16)$$

Table 3 shows the diagnostic test evaluation of PID data set. Comparison of the accuracy of the proposed framework and some methods in the literature for all clinical data set is shown in Table 4.

Inferences from the findings: The gestational diabetes 3D surface output was computed for PID dataset. As a 3D surface plot, each point corresponds to a specific input and output feature margin value. From the 3D outcome of PID dataset the clinical inference obtained for diagnosing gestational diabetes is presented in Table 5.

If input feature number of times pregnant value increases from 1 and the plasma glucose level is >153.3 mg dL^{-1} and diastolic blood pressure level is >77.86 mmHg and triceps skin fold thickness is >7mm and

serum insulin level is below 133 muU mL^{-1} and BMI greater than 25.43 kg/m^2 and diabetes pedigree function is less than 0.760 age is greater than 23.29 years then the patient have Gestational diabetes.

CONCLUSION

The experimental results show that the CDSS proposed in this research provides high classification accuracy and outperforms the other algorithms in the diagnosis of GDM. Though the fuzzification approach and the classification framework used in this CDSS is tailored for diabetes dataset, they can be also adopted for other clinical datasets. Use of metaheuristic approaches instead of exact interpolation property of RBFNN for adjusting the parameters of the ANN may yield novel and enhanced results. Moreover, the CDSS can be used for diagnosis of other diseases to promote the welfare of health services.

REFERENCES

American Diabetes Association, 2014. Diagnosis and classification of diabetes mellitus. Diabetes Care, 37: S81-S90.

Beloufa, F. and M.A. Chikh, 2013. Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm. Comput. Methods Programs Biomed., 112: 92-103.

Blake, C.L. and C.J. Merz, 1998. UCI Repository of Machine Learning Databases. 1st Edn., University of California, Irvine, CA.

Broomhead, D.S. and D. Lowe, 1988. Radial basis functions, multi-variable functional interpolation and adaptive networks (No. RSRE-MEMO-4148). MCS Thesis, Royal Signals And Radar Establishment Malvern, UK.

Cetisli, B., 2010. Development of an adaptive neuro-fuzzy classifier using linguistic hedges: Part 1. Expert Syst. Appl., 37: 6093-6101.

Christopher, J.J., H.K. Nehemiah and A. Kannan, 2015. A clinical decision support system for diagnosis of allergic rhinitis based on intradermal skin tests. Comput. Biol. Med., 65: 76-84.

Dennis, B. and S. Muthukrishnan, 2014. AGFS: adaptive genetic fuzzy system for medical data classification. Appl. Soft Comput., 25: 242-252.

Elgawi, O.H. and O. Hasegawa, 2007. Online incremental random forests. Proceeding of the International Conference on Machine Vision, ICMV 2007, December 28-29, 2007, IEEE, Tokyo, Japan, ISBN:978-1-4244-1624-0, pp: 102-106.

Elizabeth, D.S., A. Kannan and H.K. Nehemiah, 2009. Computer-aided diagnosis system for the detection of bronchiectasis in chest computed tomography images. Intl. J. Imaging Syst. Technol., 19: 290-298.

Ephzibah, E.P., 2011. Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis. Intl. J. Soft Comput., 2: 1-10.

Ganji, M.F. and M.S. Abadeh, 2011. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Syst. Appl., 38: 14650-14659.

Ghazavi, S.N. and T.W. Liao, 2008. Medical data mining by fuzzy modeling with selected features. Artif. Intell. Med., 43: 195-206.

Ghosh, S., S. Biswas, D. Sarkar and P.P. Sarkar, 2014. A novel neuro-fuzzy classification technique for data mining. Egypt. Inf. J., 15: 129-147.

Goncalves, L.B., M.M.B.R. Vellasco, M.A.C. Pacheco and D.F.J. Souza, 2006. Inverted hierarchical neuro-fuzzy BSP system: A novel neuro-fuzzy model for pattern classification and rule extraction in databases. IEEE Trans. Syst. Man Cybern., 36: 236-248.

Guzaitis, J., A. Verikas, A. Gelzinis and M. Bacauskiene, 2009. A Framework for Designing a Fuzzy Rule-Based Classifier. In: Algorithmic Decision Theory, Francesca, R. and T. Alexis (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-04427-4, pp: 434-445.

Jaganathan, P., K. Thangavel, A. Pethalakshmi and M. Karnan, 2007. Classification rule discovery with ant colony optimization and improved quick reduct algorithm. IAENG. Int. J. Comput. Sci., 33: 50-55.

Jane, N.Y., K.H. Nehemiah and K. Arputharaj, 2016. A temporal mining framework for classifying un-evenly spaced clinical data. Appl. Clin. Inf., 7: 1-21.

Kahramanli, H. and N. Allahverdi, 2008. Design of a hybrid system for the Diabetes and heart diseases. Expert Syst. Appl., 35: 82-89.

Lee, C.S. and M.H. Wang, 2011. A fuzzy expert system for diabetes decision support application. IEEE. Trans. Syst. Man Cybern. Part B Cybern., 41: 139-153.

Lekkas, S. and L. Mikhailov, 2010. Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. Artif. Intell. Med., 50: 117-126.

Luukka, P. and T. Leppalampi, 2006. Similarity classifier with generalized mean applied to medical data. Comput. Biol. Med., 36: 1026-1040.

Mohamadi, H., J. Habibi, M.S. Abadeh and H. Saadi, 2008. Data mining with a simulated annealing based fuzzy classification system. Pattern Recognit., 41: 1824-1833.

Nehemiah, H.K. and A. Kannan, 2006. A diagnostic decision support system for adverse drug reaction using temporal reasoning. Int. J. Artif. Intell. Mach. Learn., 6: 79-86.

Park, M.S. and J.Y. Choi, 2008. Novel Incremental Principal Component Analysis with Improved Performance. In: Structural, Syntactic and Statistical Pattern Recognition, Niels, D.V.L., T. Kasparis, F. Roli, J.T. Kwok and G.C. Anagnostopoulos *et al.* (Eds.). Springer, Berlin, Heidelberg, ISBN:978-3-540-89688-3, pp: 592-601.

Polat, K. and S. Gunes, 2007a. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Process., 17: 702-710.

Polat, K. and S. Gunes, 2007b. An improved approach to medical data sets classification: artificial immune recognition system with fuzzy resource allocation mechanism. Exp. Syst., 24: 252-270.

Pulkkinen, P. and H. Koivisto, 2008. Fuzzy classifier identification using decision tree and multiobjective evolutionary algorithms. Int. J. Approximate Reasoning, 48: 526-543.

Quteishat, A., C.P. Lim and K.S. Tan, 2010. A modified fuzzy min-max neural network with a genetic-algorithm-based rule extractor for pattern classification. IEEE Trans. Man Cybern. Syst. Hum., 40: 641-650.

Sahan, S., K. Polat, H. Kodaz and S. Gunes, 2005. The Medical Applications of Attribute Weighted Artificial Immune System Diagnosis of Heart and Diabetes Diseases. In: Artificial Immune Systems, Christian, J., M.L. Pilat, P.J. Bentley and J.I. Timmis (Eds.). Springer, Berlin, Heidelberg, ISBN: 978-3-540-28175-7, pp: 456-468.

Shortliffe, E.H., R. Davis, S.G. Axline, B.G. Buchanan and C.C. Green *et al.*, 1975. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. Comput. Biomed. Res., 8: 303-320.

Tang, P.H. and M.H. Tseng, 2009. Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification. Proceeding of the 2009 International Conference on Machine Learning and Cybernetics, July 12-15, 2009, IEEE, Taiwan, China, ISBN:978-1-4244-3702-3, pp: 3070-3075.

Temurtas, H., N. Yumusak and F. Temurtas, 2009. A comparative study on diabetes disease diagnosis using neural networks. Expert Syst. Appl., 36: 8610-8615.

Witten, I.H. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufman, San Francisco, CA., USA., ISBN-13: 9780080477022, Pages: 560.