J Med Syst (2017) 41:178
DOI 10.1007/s10916-017-0823-3

PATIENT FACING SYSTEMS

Evolutionary and Neural Computing Based Decision Support System for Disease Diagnosis from Clinical Data Sets in Medical Practice

M. Sudha [1]

Received: 17 August 2017 / Accepted: 18 September 2017

Abstract As a recent trend, various computational intelligence and machine learning approaches have been used for mining inferences hidden in the large clinical databases to assist the clinician in strategic decision making. In any target data the irrelevant information may be detrimental, causing confusion for the mining algorithm and degrades the prediction outcome. To address this issue, this study attempts to identify an intelligent approach to assist disease diagnostic procedure using an optimal set of attributes instead of all attributes present in the clinical data set. In this proposed Application Specific Intelligent Computing (ASIC) decision support system, a rough set based genetic algorithm is employed in pre-processing phase and a back propagation neural network is applied in training and testing phase. ASIC has two phases, the first phase handles outliers, noisy data, and missing values to obtain a qualitative target data to generate appropriate attribute reduct sets from the input data using rough computing based genetic algorithm centred on a relative fitness function measure. The succeeding phase of this system involves both training and testing of back propagation neural network classifier on the selected reducts. The model performance is evaluated with widely adopted existing classifiers. The proposed ASIC system for clinical decision support has been tested with breast cancer, fertility diagnosis and heart disease data set from the University of California at Irvine (UCI) machine learning repository. The proposed system outperformed the existing approaches attaining the accuracy rate of 95.33%, 97.61%, and 93.04% for breast cancer, fertility issue and heart disease diagnosis.

Keywords Clinical decision support · Disease prediction · Neural network · Genetic algorithm · Rough set · Feature reduction and hybrid computing

Introduction

Disease prediction involves relatively lively, multifaceted, real-time, and unexpected characteristics. Recently the demand for data-driven predictive approaches to improve the forecasting accuracy has heightened to facilitate interactive, user-friendly tools. The majority of work reported for medical diagnosis problems has been done with an expert consultant or physician. Therefore a more attention is required towards developing a soft computing approach for decision support in disease diagnosis from the clinical data set. Information in clinical databases surpassed the human competencies to distinguish and identify the hidden knowledge. This current challenge has raised a demand for new computational intelligence approaches. It has witnessed that existing data mining techniques often did not match the current requirements. Therefore, this investigation has examined the feasibility of using rough set computing coupled with an evolutionary computing approach to fill the existing gap. The outcome or the decision, especially disease diagnosis, is strongly dependent on several chaotic factors. It is important to take all significant factors into consideration for disease diagnosis. The prediction may change considerably with very small changes in the input parameters. Also in much clinical data set or patient records, the data quality is often not very good. The primary focus of this research is to investigate the existing methods and to develop a reliable intelligent system incorporating

This article is part of the Topical Collection on *Patient Facing Systems*

✉ M. Sudha
msudha@vit.ac.in

[1] School of Information Technology and Engineering, VIT University, Vellore, India

Published online: 27 September 2017

rough set, evolutionary computing, and neural network entitled as application specific intelligent computing systems in clinical decision support.

This article is organized as follows. "Background and Related works section", outlines the background of rough computing, neural networks and related works on diseases diagnostic systems based on computational intelligence approaches. The methodology and the clinical dataset used for this study are presented in "Related works, Materials and methods and Data set description section". Next, in the "ASIC system framework section", the proposed computational intelligence based medical diagnostic process is described. Finally, the complete structure of ASIC and the experimental results on medical data sets for the diagnosis of breast cancer, semen quality analysis (fertility diagnosis) and heart disease is reported in "Results and discussions section".

Background

Rough computing approach as introduced by Pawlak is a mathematical model that deals with imperfect knowledge. It requires no prior knowledge of the target data set. Indiscernible relation is the basis of the rough set theory and according to that every object in the universal-set and objects characterized by same values are indiscernible and objects with different values are discernible [1]. Rough set approaches widely applied in the field of intelligent data analysis. It is especially suitable for parallel processing, finding minimal data sets, and supplying practical techniques to find hidden knowledge in the database. Rough computing is used for producing decision rule set of information, and the results obtained can be interpreted clearly. [2] described the mathematical viewpoint of the rough set theory and applicability of rough set theory for rule discovery from a decision table. [3] Introduced two attribute reduction algorithms namely Johnson's reduct and the object reduct using feature weighting technique. Mostly the choice of algorithm used may be dependent on the data and its attributes [2]. Yao and Zhao defined a reduct method based on discernibility matrix simplification [4].

In [5], various applications of rough set concept for handling heterogeneous data involving different data types with imperfect knowledge was reported by Yaho and Zhao. The indiscernibility is the mathematical basis of rough set theory Pawlak [6]. In [7], a scatter- search-rough set attribute reduction was introduced and the effect of this approach has been widely studied and tested using some well-known data sets from the University of California, Irvine (UCI) machine learning repository. A feature selection approach integrated with the bee colony was proposed, this method produced minimal reducts for medical data sets [8].

After pre-processing the raw dataset using rough computing, the target data has to undergo training using a machine learning approach to predict the hidden inference. Therefore this research applies a proven neural computing approach to enhance the machine learning process. Werbos introduced the Back Propagation Neural Network (BPNN) in 1974 [9] and then popularized by Rumelhart [10]. The adaptability or learning potential of techniques makes it very useful in problem-solving for highly nonlinear phenomena [11]. Artificial Neural Network (ANN) methods, bayesian network and other data mining approaches are applied for the diagnosis of thyroid disease, thyroid malfunction and breast cancer survivability diagnosis [12–14]. Feature selection techniques based on attribute weighting using genetic algorithm, novel feature selection approaches and hybrid systems are applied in heart disease, liver disorder medical datasets classification [15–17]. A feed forward back propagation neural network method was applied for the Prediction of nephritis disease diagnosis [18]. Artificial neural network structures are suitable for chest disease diagnosis and can serve as learning based medical decision support system for assisting the consultants in their diagnosis decisions [19].

A probabilistic neural network model used for Malignant mesothelioma disease diagnosis has reported the best classification accuracy of 96.30% via three fold cross validation on the dataset prepared from new patient's hospital reports of medicine database from south east region of Turkey [20]. Probabilistic neural networks reported best accuracies on thyroid disease diagnosis [21]. Keles and Keles reinstated expert systems based on artificial intelligences as most widely used method in clinical system [22].

Related works

As a recent trend data mining and artificial intelligence approaches are widely adopted in developing clinical decision support systems [23]. Also, a recent analysis conducted on hybrid classification systems for heart disease diagnosis using relief and roughest method has reported improved classification accuracy of 92.59% for jack-knife cross validation when trained using reduct set [24]. In [25], extreme learning machine (ELM) algorithm has been applied to diagnose Erythematous Diseases. The experimental results revealed that the proposed ELM model can learn better than any other ANN approaches. In [26], a study on the diagnosis of Erythemato-Squamous diseases system based on two data mining techniques, support vector machine and ANN. The techniques were applied as a combined confidential weighted voting scheme and attained the utmost accuracy during the training and testing phases.

A relative analysis of the application of Naive Bayes, J48 decision tree induction and Multilayer Perceptron in medical diagnosis process and their experimental result indicates

Naive Bayes as a suitable classifier in diseases diagnosis or classification process [27]. In [28], a new framework that represents a comprehensive guideline for selecting suitable algorithms needed for different steps of an automatic diagnostic procedure for ensuring timely diagnosis of skin cancer has been described.

An ANN based diseases diagnosis system on predicting different types of skin diseases and accomplished an accuracy of 90% [29]. In [30] it is stated that a critical domain relevant to most of the Medical Diagnosis Decision Systems (MDDSs) is confirmation, assessment, and unending quality assurance. In MDDS, precision is being the utmost significant metric to be evaluated [31, 32].

In [33], a Multilayer Perceptron model has been applied to evaluate prediction of the seminal quality from the data of the environmental factors using back propagation algorithm. In [34], a minimum distance based K-NN classifier was incorporated to classify the patients based on the attributes collected from medical field. The experimental result reported Fuzzy K-NN classifier as a suitable tool for medical diagnosis when compared with other parametric techniques.

Materials and methods

The following tasks are identified and implemented to attain the primary intent of this investigation in modeling ASIC framework for clinical decision support:

- Collect clinical data set for medical decision support system and perform the necessary pre-processing activities for removing missing values, noisy data (data cleaning) to attain effective feature selection.
- To investigate different data-driven approaches used in modeling disease diagnosis, with a specific focus on breast cancer, fertility analysis and heart disease problems.
- Apply generic data mining approaches to different feature subsets generated using rough set based feature selection, its performance and its capability are evaluated regarding accuracy and misclassification rate.
- Develop an intelligent computing system based diagnostic system using artificial intelligence to distinguish its performance with the existing data mining methods.
- Evaluate the performance of ASIC framework on the clinical data set and compare with the existing classifiers prediction accuracy.

Data set description

For this study, three clinical data sets have been carefully chosen from the UCI machine learning repository. Those are breast cancer, fertility diagnosis and heart disease data set. The diseases diagnosis considers the specific set of attributes that describes the symptoms in the human. The metrics used in representing the attribute values during the process of breast cancer diagnosis with respect to the classification method is projected in Table 1. The classifier states whether the patient's breast tissue is malignant or benign. The fertility diagnosis process involves checking the semen quality of a male and the ASIC framework is subject to predict the quality based on the specified attributes.

The attributes and the domain values used in the evaluation of fertility diagnosis are shown in Table 2. It consists of '9' attributes which are subject to feature selection process which selects the most significant parameters to conduct the prediction or classification process. The heart disease diagnosis process makes use of about '13' attributes and the domain values of each attribute and the type of the values are as in Table 3. The target data sets used for training and testing the classification are reduct sets that consists of significant features or attributes. This attribute selection is performed in the preprocessing phase of this ASIC system. The attribute information related to breast cancer, semen quality diagnosis and heart disease are projected in Tables 1, 2 and 3. The breast cancer, semen quality diagnosis and heart disease attributes are labelled as (b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8 and b_9); (s_1, s_2, s_3, s_4, s_5, s6, s_7, s_8 and s_9) and (h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8, h_9, h_{10}, h_{11}, h_{12} and h_{13}) for representation simplicity.

ASIC system framework

ASIC is designed mainly based on artificial neural network architecture. The ANN model applied is designed in the form of a multi-layer perceptron with sigmoid non-linearity activation function. This adaptive rough evolutionary neuro approach abridged as ASIC comprises an effective feature selection approach based on rough and genetic computing

Table 1 Attribute information of Breast Cancer data

Attribute name	Domain	Missing value
Clump thickness (b_1)	1–10	0
Uniformity of cell size(b_2)	1–10	0
Uniformity of cell shape (b_3)	1–10	0
Marginal adhesion (b_4)	1–10	0
Single epithelial cell size (b_5)	1–10	17
Bare nucleoli (b_6)	1–10	0
Bland chromatin (b_7)	1–10	0
Normal nucleoli (b_8)	1–10	0
Mitosis (b_9)	1–10	0
Class	2 for benign, 4 for malignant	0

Source- UCI repository Irvine

Table 2 Attribute information of Fertility data

Attribute name	Domain	Missing value
Season (s_1)	1, −0.33, 0.33, 1	0
Age (s_2)	0–1	0
Childish diseases (s_3)	0–1	0
Accident or Serious trauma (s_4)	0–1	0
Surgical intervention (s_5)	0–1	17
High fevers (s_6)	0–1	0
Frequency of alcohol consumption (s_7)	0–1	0
smoking habit (s_8)	0–1	0
Sitting hours per day (s_9)	0–1	0
Class (Diagnosis result)	1 for normal, 0 for altered	0

Source - UCI repository Irvine

techniques. The proposed hybrid intelligent systems framework differs from other existing models by coupling three new-fangled blends of intelligent techniques. The benefits of rough set and the genetic algorithm approach could identify a globally optimal input for training the network. The input data can be one of the factors that may influence the output of the architecture. Subsequently, a back propagation algorithm is employed to train neural network architecture [35].

ASIC pre-processing phase

The pre-processing phase identifies the missing values, noisy data present in the raw data set. Initially, rough set based feature selection process generates the reduct sets based on discernibility matrix approach. The subsequent process of identifying the most significant feature reduct or attribute subset is attained by means of the proposed Application Specific reduct-selection using Genetic Algorithm (ARGA), based on novel fitness and the relative fitness functions are determined using eq. 1 and eq. 2.

Rough computing adopted as input data selection method identifies the possible 'n' reducts set or feature subsets for the complete feature set. In this ASIC's pre-processing phase initial set of reduct generation is accomplished using Rough Set Exploration system RSES 2.2 [36] and consequently the

Table 3 Attribute information of Heart disease data

Attribute name	Data type	Domain
Age (Patient age in year) (h_1)	Numerical	29 to 77
Sex (Gender) (h_2)	Binary	0 = female
		1 = male
Chp (Chest pain type) (h_3)	Nominal	1 = typical angina,
		2 = atypical angina
		3 = nonanginal pain,
		4 = asymptomatic
Bp (Resting blood pressure) (h_4)	Numerical	94 to 200
Sch (Serum cholesterol) (h_5)	Numerical	126 to 564
Fbs (Fasting blood sugar >120 mg/dL) (h_6)	Binary	0 = false
		1 = True
Ecg (Resting electrocardiographic result) (h_7)	Nominal	0 = normal
		1 = having ST-T wave abnormality
		2 = left ventricular hypertrophy
Mhrt (Maximum heart rate) (h_8)	Numerical	71 to 200
Exian (Exercise induced angina) (h_9)	Binary	0 = no
		1 = yes
Opk (Old peak) (h_{10})	Numerical	Continuous (0 to 6.2)
Slope (Slope of peak exercise ST segment) (h_{11})	Nominal	1 = upsloping
		2 = flat
		3 = downsloping
Vessel (Number of major vessels) (h_{12})	Nominal	0 to 3
Thal (Defect type) (h_{13})	Nominal	3 = normal, 6 = fixed defect,
		7 = reversible defect
Class (Heart disease)	Binary	0 = absence,
		1 = presence

Source - UCI repository Irvine

optimal reduct subset selection is achieved using the proposed ARGA. The breast cancer, fertility diagnosis and heart disease dataset subject to ASIC pre-processing stage has set of conditional attributes and a decision attribute. The breast cancer data set consists of nine conditional attributes and one decision (class) attribute. The decision attribute is a binary class variable, it exhibits two decision values namely benign (non-cancerous tumour) and malignant (cancerous tumour) as described in Table 1. In case of fertility and heart disease dataset there are nine and thirteen conditional attributes and one decision attribute representing two possible class variables.

Application Specific reducts selection using Genetic Algorithm

1. Begin
2. Let {GFs-$_{Rd}$} be the entire set of reduct sets
3. S = (Sum of all reducts is represented as {GFs-$_{Rd}$}
4. Encode the input reduct data set in bit string format
5. {
   a. Initialize the Initial population
   b. P = n
   c. Compute the fitness function $F_{fn}(X)$ of a reduct set
   d. Estimate the relative Fitness Function $RF_{fn}(x)$
   e. If $RF_n(x) \geq \text{Avg}RF_n(x)$ then include the set in to new population
   f. Apply a single point crossover function
   g. Implement mutation operation
   h. Else ignore
   i. P-
6. }
7. Repeat step 4 to 6 until a desired stopping criterion or s =0
8. Return {RFs-$_{Red}$}
9. Terminate

The complete feature set represented as {b1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9} corresponds to the breast cancer attributes, The set of all reducts generated using RSES 2.2 software is represented in a bit-string format as shown in Table 5 for optimal reduct selection process. From among the reducts generated for breast cancer dataset, the sample reduct set {b_1, b_3, b_5, b_6, b_7} generated using RSES 2.2 is represented as {1, 0, 1, 0, 1, 1, 1, 0, 0} which is the chromosomal representation of the sample reduct set. That is, the attributes present in the reduct set is represented as '1' and least significant attributes that are removed from the complete set is notated as '0', this the targeted bit string format input for ARGA.

The reduct sets that are subsequently encoded to bit string format represents the chromosomal format required by the GA in ARGA approach for optimal reduct selection based on the relative fitness function as given in eq. 2. Correspondingly the same procedure is applied on all the reducts sets obtained for breast cancer, semen quality analysis and heart disease diagnosis data set using RSES 2.2. The input reducts dependent and vary according to the raw input data. From among the reducts appropriate reduct set is identified using ARGA as shown in Fig. 1 and 2.The input to ARGA is a reduct set in the form of bit string format.

$$\text{Fitnessfunction } F(x) = C_{n1}/\text{Bit\_str}_{len} \quad (1)$$

$$\text{RelativeFitnessfunction } RF_{fn}(x) = F_{fn}(x)/\text{Avg } F_{fn}(x) \quad (2)$$

Where, C_{n1} is the total count of '1' bit value in each chromosome, Bit_str_{len} is the total number of bits in the chromosome or bit string length it is constant value of the sample data set and Avg F(x) is the average of the fitness function.

Breast cancer: reduct, {1, 0, 1, 0, 1, 1, 1, 0, 0} and the corresponding Cn count and the $\text{Bit\_str}_{len\ is}$ [('Cn1' = 5) and (Bit_str_{len} = 9)].

Correspondingly, the complete feature set of fertility and heart disease data set is (s_1, s_2, s_3, s_4, s_5, s6, s_7, s_8 and s_9) and (h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8, h_9, h_{10}, h_{11}, h_{12} and h_{13}). The $\text{Bit\_str}_{len\ is}$ the count of the all features in the sample data. Therefore, fertility diagnosis [Bit_str_{len} = 9)] and for heart disease diagnosis [Bit_str_{len} = 13)]. Similarly the 'Cn1' value is estimated for all possible reduct sets for breast cancer, fertility diagnosis and heart diseases dataset in turn to determine the relative fitness function. The optimal feature set determined based on ARGA is used for training the neural computing based classification model. The optimal reducts obtained using the proposed ARGA in bit string format is projected in Table 4 and the corresponding reducts are shown in Table 5.

The optimal reducts obtained for breast cancer data is shown in Table 5. Similarly the reduct sets generated for semen quality analysis and heart disease using RSES tool is applied as target input to ARGA to identify the optimal reducts. These optimal reduct set obtained using ARGA is used for training the classifiers to predict the diseases and in turn enable us to identify the significant symptoms for strategic decision making in the next phase.

ASIC – Training and testing phase

The ASIC systems training and testing process is performed using back propagation neural network architecture. In back propagation neural network's forwarding phase, the input form is applied, and the desired output is measured. The estimated error of each neuron is the differentiation between the targeted and the output attained. This projected error is consequently used to modify or adjust the weights in order to minimize the

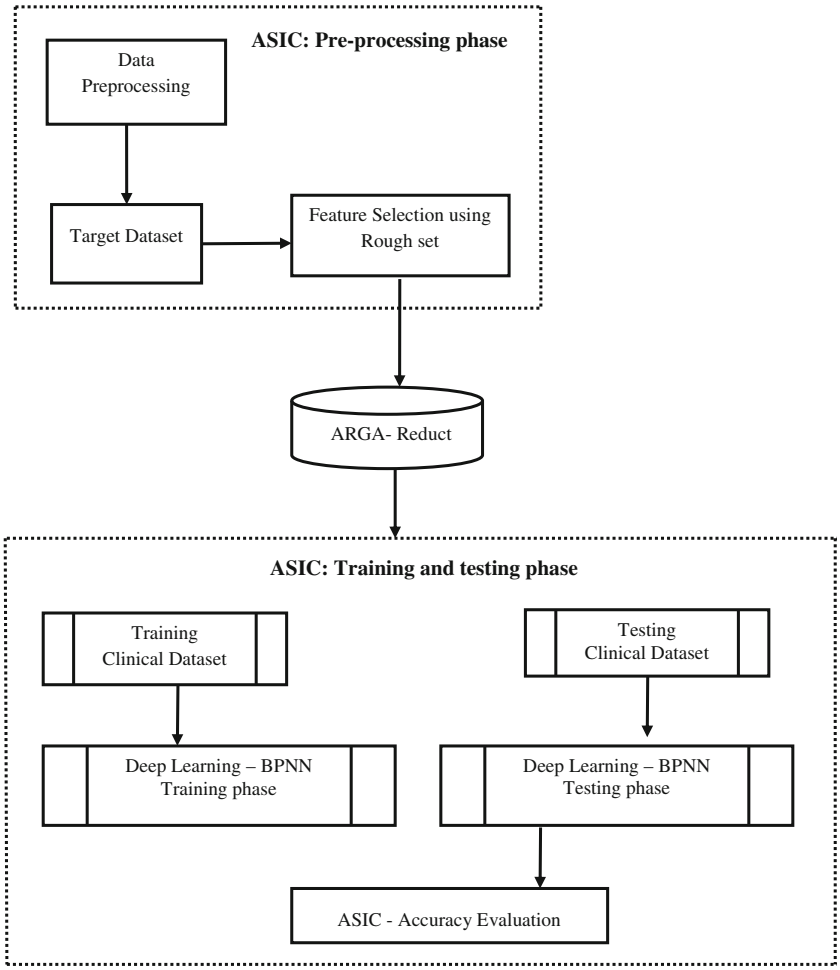

Fig. 1 ASIC system for clinical decision support

error to achieve the desired accuracy. This process is repeated until the desired output is achieved with nominal error. The proposed BPNN based ASIC is implemented in dot Net environment.

Optimal reducts are set as input to the n-nodes of input layer. The units of input layer have to forward this information to all the nodes in the hidden layer. At any hidden node (h), the data received from all the input nodes and the bias mode of the input layer are summed up as $I_1 + I_2 + I_3 + ... I_n + \text{bias}$. The proposed design implements sigmoid activation function and the single neuron in output layer of BPNN outputs the desired accuracy. Initially, BPNN weights are set to some small random numbers between (−1 and +1) or (−0.5 and +0.5).

In forwarding phase, the input pattern is applied, and the desired output is estimated. The estimated error of each neuron is the differentiation between the targeted and the attained accuracy. This estimated error is subsequently used to modify or adjust the weights in order to minimize the error to achieve the desired optimal accuracy. Next, the Output of each neuron is tuned to get closer to its target on the reverse pass. This process is repeated until the desired output is achieved with nominal error.

Back propagation neural network algorithm

1. Begin with randomly chosen weights;
2. While MSE is above desired threshold and computational bounds are not exceeded, do
3. {
4. For each input pattern x_p, $1 \le p \le P$,
5. {
   a. Compute hidden node inputs;
   b. Compute hidden node outputs;
   c. Compute inputs to the output nodes;
   d. Compute the network outputs;
   e. Compute the error between output and desired output;
   f. Modify the weights between hidden and output nodes;
   g. Modify the weights between input and hidden nodes;
   h. }
6. }
7. End-for
8. End-while

Related to this proposed approach, prediction of heart disease using three classification techniques namely

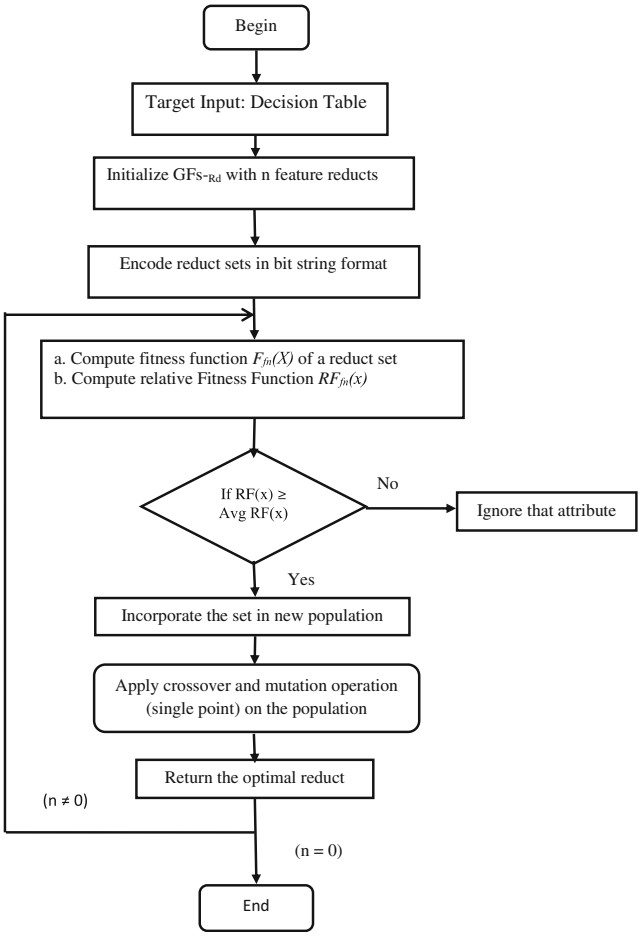

Fig. 2 ARGA process flow diagram

Decision Trees, naive bayes and K-nearest neighbour has reported these classification techniques as significant tool for medical diagnosis [37]. Similarly, a new algorithm assimilating naive bayes with genetic algorithm has been used for heart disease diagnosis [38]. Classification tree algorithms are employed in risk assessment of heart failure rate [39]. Correspondingly, support vector machine is used in Parkinson's disease prediction [40]. As an advancement of several existing work the proposed method stabs to identify the significant attribute set influencing the classifier performance.

Table 4 Reduct in bit string notation

b1	b2	b3	b4	b5	b6	b7	b8	b9	F(x)	RF(x)	Positive Region
1	1	1	1	1	0	1	0	0	0.75	1.25	0.9916
1	1	1	0	1	1	1	0	0	0.75	1.25	0.9932
1	0	1	1	1	1	0	0	0	0.625	1.04	0.9679
1	0	1	0	1	1	1	0	0	0.75	1.25	0.9981
0	1	1	1	1	0	1	0	0	0.76	1.24	0.9963
0	0	1	1	1	1	1	0	0	0.76	1.24	0.9765

Table 5 Optimal reduct sets generated using ARGA

Reduct	Selected reduct set
Optimal reduct (1)	{b1, b2, b3, b4, b5, b7}
Optimal reduct (2)	{b1, b2, b3, b5, b6, b7}
Optimal reduct (3)	{b1, b3, b4, b5, b6, b7}
Optimal reduct (4)	{b1, b3, b5, b6, b7}
Optimal reduct (5)	{b2, b3, b4, b5, b7}
Optimal reduct (6)	{b3, b4, b5, b6, b7}

Therefore, this work attempt to train the widely adopted existing data mining techniques such as naive bayes, classification tree and back propagation classifiers and the proposed ASIC in predicting the breast cancer, fertility issues and heart disease.

Results and discussions

The precision rate is the percent of instances that are appropriately classified by the predictive model for the specified testing and training set by means of ten-fold cross validation in WEKA [41]. The model accuracy is calculated as the average error across the ten folds. The key point of this cross-validation is that it uses every possible sample for testing, and it can avoid an ill-fated split. The confusion matrix is an expedient tool for examining how well a classifier can identify instances of different classes [35]. Tp and Tn tell us when the classifier attainment is accurate, while FP and FN covey us the classifier has wrongly identified the rows or instance. The prediction accuracy, RMSE value and the misclassification rate is determined from the confusion matrix of each classifier during its training and testing phase using WEKA tool.

A confusion matrix is expected as the most relevant measure for 10-fold cross-validation. A confusion matrix describes the actual and predicted classification done by each classifier individually.

$$\text{Accuracy} \quad \text{Rate}\left(\text{Ac}^{\text{R}}\right) = \frac{\text{Tp} + \text{Tn}}{(\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn})} \tag{3}$$

$$\text{MisclassificationRate (McR) orErrorRate} = \frac{\text{Fp} + \text{Fn}}{(\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn})} \tag{4}$$

Where, Tp - True positive, Tn - True negative, Fp - False positive and Fn - False negative.

The false positive rate (Fp) is the percent of negative cases that were erroneously classified as positive. The true negative rate (Tn) is defined as the amount of negative

Table 6 Accuracy (%) of proposed ASIC Vs Existing Classifiers

Classifier - Model	Breast Cancer	Fertility diagnosis	Heart disease
Random Forest	82.27	81.27	81.27
Bayesian Network	84.17	79.17	79.07
Decision Tree	84.27	88.07	81.07
Voted Perceptorn	85.82	82.82	84.82
Naive Bayes	77.89	78.89	78.79
Radial Basis Function	85.34	82.34	85.24
Proposed ASIC - BPNN	95.31	97.61	93.04

cases that are classified appropriately. The false negative rate (Fn) is the percent of positive cases that were mistakenly classified as negative. The accuracy rate is the percent of instances that are correctly classified by the classifier for the specified test set is shown in Table 6. The root mean squared error (RMSE) is shown in Table 7. The incorrectly classified instances determine the error rate or misclassification rate of the classifier is projected in Table 8.The learning models are evaluated based on the measures in eq. (3 to 4).

The accuracy (Ac^R) is the percent of the sum of predictions that were true. It is estimated using the eq. (3). Table 6, presents the accuracy attained by the proposed backpropagation neural network (BPNN approach incorporated in ASIC) and existing classification techniques for breast cancer, fertility and heart disease diagnosis. RMSE represents the sample standard deviation of various classifiers and the differences between predicted and observed values for the same sample dataset. Table 7, presents the root mean squared error reported by the proposed backpropagation neural network (BPNN approach incorporated in ASIC) and existing classification techniques for disease breast cancer, fertility and heart disease diagnosis.

Table 7 RMSE of proposed ASIC Vs Existing Classifiers

Classifier - Model	Breast Cancer	Fertility diagnosis	Heart disease
Random Forest	0.38	0.38	0.38
Bayesian Network	0.41	0.41	0.41
Decision Tree	0.38	0.38	0.38
Voted Perceptorn	0.32	0.32	0.32
Naive Bayes	0.36	0.36	0.36
Radial Basis Function	0.34	0.34	0.34
Proposed ASIC - BPNN	0.31	0.31	0.34

Table 8 Misclassification rate of proposed ASIC Vs Existing Classifiers

Classifier - Model	Breast Cancer	Fertility diagnosis	Heart disease
Random Forest	0.1773	0.1873	0.1873
Bayesian Network	0.1583	0.2083	0.2093
Decision Tree	0.1573	0.1193	0.1893
Voted Perceptorn	0.1418	0.1718	0.1520
Naive Bayes	0.2211	0.2111	0.2121
Radial Basis Function	0.1466	0.1766	0.1476
Proposed ASIC - BPNN	0.0469	0.0233	0.0696

Table 8, presents the percentage of wrongly classified instances or misclassification rate of the proposed backpropagation neural network (BPNN approach incorporated in ASIC) and existing classification techniques for disease breast cancer, fertility and heart disease diagnosis. The error rate of a classifier 'C' is (1- accuracy rate C) and it is also computed as in eq. 4. All the classifiers projected in the Tables 6, 7 and 8 make use of same sample dataset for this investigation. The breast cancer, fertility (semen quality analysis) diagnosis and heart disease sample dataset used for the purpose of investigating this clinical decision support system are obtained from the University of California at Irvine (UCI) machine learning repository. The accuracy attained by existing and proposed ASIC model is visualized in Fig. 3.These types of decision support systems would certainly have demanding application in future medical practices to avoid late detection and loss money if the patient has no negative result (no disease). Thus these expert decision support systems can be presented as an additional service to existing policies of integrated care for chronic-disease management [42].

The research conducted by Chaurasia and Pal to predict breast cancer using rep tree and radial basis function (RBF) network reported 74.5% accuracy and their studies stated that this model can produce fast automatic diagnostic results for other diseases [43]. Wherein the existing classifiers and the proposed system outperformed the work reported previously.The comparative analysis of various classification algorithms on diagnosis of type 2 diabetes in Iran conducted by Heydair et al. reported accuracy rate of 81.19%,95.03%, 90.85%, and 91.60% for support vector machine, decision tree, 5-nearest neighbor, and Bayesian network correspondingly. The results of their simulations revealed that the efficiency of classification techniques depends on the dataset used [44]. Similarly a hybrid prediction model attained classification accuracy of 92.38% for type-2 diabetic patient's diagnosis. The performance of the proposed method was

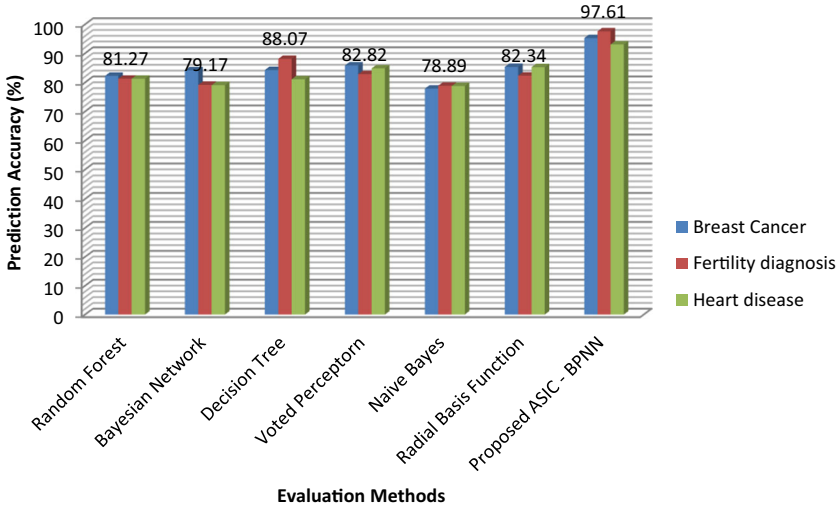

Fig. 3 Chart representation of prediction accuracy attained by proposed ASIC Vs Existing Classifiers

evaluated based on sensitivity and specificity performance measures [45].

Conclusion

The computational intelligence approaches have been applied in the medical field, which incorporated both knowledge representation tools and algorithms that perform the diagnosis. This study presents the analysis of the results obtained from ASIC focusing on the metrics that estimate the overall accuracy, root mean square error, misclassification rate, specificity and sensitivity of the system with other existing data mining techniques. Among the existing classifiers probabilistic based naive bayes approach has reported low accuracy rate on all the clinical data set of this study. Similarly bayesian network model attained low accuracy on semen quality analysis and heart disease diagnosis. The experimental result shows that the reduct set $\{b_1, b3, b_5, b_6, b_7\}$, $\{s_2, s_3, s_4, s_6, s_7, s_9\}$ and $\{h_1, h_3, h_5, h_6, h_7, h_{10}, h_{11}\}$ of breast cancer, semen quality analysis and heart disease achieves improved classification accuracy. The outcome reveals the proposed ASIC system superior to existing data mining methods.

Therefore these observational results conclude that assimilating one or more intelligent techniques (hybrid computing) may report better outcome than the generic learning algorithms. Artificial neural network based ASIC with an accuracy rate of 97.61% model has outperformed with semen quality analysis data set. As quantified before, the results confirm that the system would be a good acquaintance for specialists, where the system can suggest diagnoses and the therapeutic specialists can confirm the same to the patients.

Funding No funding for this research work.

Compliance with Ethical Standards

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Pawlak, Z., Rough sets. *International Journal of Computer and Information Sciences.* 11(5):341–356, 1982.
2. Bal, M., Rough sets theory as symbolic data mining method: an application on complete decision table. *Information Science Letters.* 2(1):35–47, 2013.
3. Ali-Khashashneh, E. A. and Q. A. Al-Radaideh (2013) Evaluation of discernibility matrix based reduct computation techniques. 5th International Conference on Computer Science and Information Technology - IEEE, Amman. 76–81
4. Yao, Y., and Zhao, Y., Discernibility matrix simplification for constructing attribute reducts. *Information Sciences.* 179(5):867–882, 2009.
5. Wei, W., Liang, J., and Qian, Y., A comparative study of rough sets for hybrid data. *Information Sciences.* 190:1–16, 2012.
6. Pawlak, Z., Rough set approach to knowledge-based decision support. *European Journal of Operational Research*. 99:48–57, 1997.
7. Wang, J., Zhang, Q., Abdel-Rahman, H., and Abdel-Monem, M.I., A rough set approach to feature selection based on scatter search metaheuristic. *Journal of Systems Science and Complexity.* 27(1): 157–168, 2014.
8. Suguna, N., and Thanushkodi, K.G., An independent rough set approach hybrid with artificial bee colony algorithm for dimensionality reduction. *American Journal of Applied Sciences.* 8(3):261–266, 2011.
9. Werbos, P (1974) Beyond regression new tools for prediction and analysis on the behavior sciences. PhD Thesis. Harvard University
10. Rumelhart, D.E., Hinton, G.E., and Williams, R.J., Learning representations by back propagating errors. *Nature.* 323:533–536, 1986.
11. Hayati, M., and Mohebi, Z., Application of artificial neural networks for temperature forecasting, World Academy of Science. *Engineering and Technology.* 1(4):654–658, 2007.

12. Ozyılmaz, L., and Yıldırım, T., Diagnosis of thyroid disease using artificial neural network methods. In: *Proceedings of ICONIP'02 nineth international conference on neural information processing*. Orchid Country Club, Singapore, pp. 2033–2036, 2002.
13. Hoshi, K., Kawakami, J., Kumagai, M., Kasahara, S., Nisimura, N., Nakamura, H., et al., An analysis of thyroid function diagnosis using Bayesian-type and SOM-type neural networks. *Chemical and Pharmaceutical Bulletin*. 53:1570–1574, 2005.
14. Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine Artificial Intelligence in Medicine*. 34(2):113–127, 2005.
15. Ozsen, S., and Gunes, S., Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. *Expert Systems with Applications.* 36(1):386–392, 2009.
16. Polat, K., and Gunes, S., A new feature selection method on classification of medical datasets: kernel F-score feature selection. *Expert Systems with Applications*. 36(7):10367–10373, 2009.
17. Kahramanli, H., and Allahverdi, N., Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*. 35(1–2):82–89, 2008.
18. Al-Shayea, Q.K., Artificial neural networks in medical diagnosis. *International Journal of Computer Science Issues.* 8(2): 150–154, 2011.
19. Er, O., Yumusak, N., and Temurtas, F., Chest diseases diagnosis using artificial neural networks. *Expert Systems with Application.* 37(12):7648–7655, 2010.
20. Er, O., Tanrikulu, A.C., Abakay, A., and Temurtas, F., An approach based on probabilistic neural network for diagnosis of Mesothelioma's disease. *Computers & Electrical Engineering*. 38(1):75–81, 2012.
21. Temurtas, F., A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications*. 36(1): 944–949, 2009.
22. Keles, A., and Keles, A., ESTDD: Expert system for thyroid diseases diagnosis. *Expert Systems with Applications.* 34(1):242–246, 2008.
23. Sudha, M., Disease diagnosis using association rule mining based knowledge inference system. *International Journal on Pharmacy and Technology.* 8(3):16369–16379, 2017.
24. Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q (2017) A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method, Computational and mathematical methods in medicine.
25. Olatunji, S.O., and Arif, H., Identification of Erythemato-Squamous skin diseases using extreme learning machine and artificial neural network. *ICTACT Journal of Soft Computing.* 4(1): 627–632, 2013.
26. Sharma, D.K., and Hota, H.S., Data mining techniques for prediction of different categories of dermatology diseases. *Journal of Management Information and Decision Sciences*. 16(2):103, 2013.
27. Danjuma K., Osofisan A O (2015) Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis. arXiv preprint arXiv:1501.00607
28. Masood A, Ali Al-Jumaily A (2013) Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. International journal of biomedical imaging.
29. Bakpo F S., Kabari L G (2011) Diagnosing Skin Diseases Using an Artificial Neural Network. In Artificial Neural Networks-Methodological Advances and Biomedical Applications. Intech.
30. Miller, R.A., Pople Jr., H.E., and Myers, J.D., Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine.* 307(8):468–476, 1982.
31. Swets, J.A., Measuring the accuracy of diagnostic systems. *Science.* 240(4857):1285–1293, 1988.
32. Huguet, J., Castineiras, M.J., and Fuentes-Arderiu, X., Diagnostic accuracy evaluation using ROC curve analysis. *Scandinavian journal of clinical and laboratory investigation.* 53(7):693–699, 1993.
33. Gil, D., Girela, J.L., De Juan, J., Gomez-Torres, M.J., and Johnsson, M., Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications.* 39(16):12564–12573, 2012.
34. Krishnaiah, V., G. Narsimha, and N. Subhash Chandra (2015) Heart disease prediction system using data mining technique by fuzzy K-NN approach, Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI)
35. Han, J., Kamber, M., and Pei, J., *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, Waltham, USA, 2011.
36. Skowron, A., J. Bazan., N. H. Son and J. Wroblewski (2005) RSES 2.2 user's guide, Institute of Mathematics, Warsaw University, Warsaw, Poland
37. Joshi, S., and Nair, M.K., Prediction of heart disease using classification based data mining techniques. *Computational Intelligence in Data Mining- Springer.* 2:503–511, 2015.
38. Kumar, S., and Sahoo, G., Classification of heart disease using Naive Bayes and genetic algorithm. *In Computational Intelligence in Data Mining- Springer.* 2:269–282, 2015.
39. Melillo, P., De Luca, N., Bracale, M., and Pecchia, L., Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE Journal of Biomedical and Health Informatics*. 17(3):727–733, 2013.
40. Yadav, G., Kumar, Y., and Sahoo, G., Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical, and support vector machine classifiers. *Indian J. Med. Sci.* 65(6):231, 2011.
41. Witten, I. H. and E. Frank (2005) Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco. P. 411
42. Velickovski, F., Ceccaroni, L., Roca, J., Burgos, F., Galdiz, J.B., Marina, N., and Lluch-Ariet, M., Clinical Decision Support Systems (CDSS) for preventive management of COPD patients. *Journal of translational medicine.* 12(2):28, 2014.
43. Chaurasia, V., and Pal, S., Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. *International Journal of Computer Science and Mobile Computing.* 3:10–22, 2014.
44. Heydari, M., Teimouri, M., Heshmati, Z., and Alavinia, S.M., Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. *International Journal of Diabetes in Developing Countries.* 36(2):167–173, 2016.
45. Patil, B.M., Joshi, R.C., and Toshniwal, D., Hybrid prediction model for type-2 diabetic patients. *Expert systems with applications*. 37(12):8102–8108, 2010.