Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 4, July 2017, pp. 348-361
DOI: 10.23940/ijpe.17.04.p2.348361

Impact of Genetic Optimization on the Prediction Performance of Case-Based Reasoning Algorithm in Liver Disease

Sakshi Takkar and Aman Singh*

Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

Abstract

Liver illness is the most hazardous ailment that influences a large number of individuals consistently and ends man's life. An effective diagnosis model is required in the process of liver disease treatment. This study accordingly aims to employ Case-Based Reasoning (CBR) methodology supported by Genetic Algorithm (GA) to optimize the prediction results of liver disease and to analyze their performances on different datasets. CBR methodology has been implemented to find the prediction results of liver disease for different datasets. We proposed a GA-based CBR framework to compare its performance with CBR in order to observe how effective it is at predicting liver illness. CBR prediction accuracy is very low so it is not very much appreciated. The proposed GA-CBR integrated model outperforms the CBR model by achieving better accuracy for all used datasets of liver disease. In this optimization of weights of features and selection of suitable instances are done simultaneously rather than separately. This leads to better prediction performance as compare to independent models. The outcome of this model illustrates that performance of usual CBR enhances fundamentally by utilizing our integrated GA-CBR model approach.

Keywords: genetic algorithms; instance selection; case-based reasoning; feature weighting; classification; liver diagnosis; integrated model.

(Submitted on March 17, 2017; First Revised on April 22, 2017; Second Revised on May 22, 2017; Accepted on May 24, 2017)

1. Introduction

In a clinical environment, medical practitioners diagnose or treat diseases based upon some physical tests or laboratory outcomes, patient's medical information and their responses to inquiries. As diagnosing the disease of a patient is a complicated task due to which several things may go wrong. The data given by the patients may encapsulate excess and noisy manifestations, especially when the patients experience the ill effects of numerous sicknesses. Sometimes, symptoms are either wrongly recognized by the physicians or poorly described by the patients to experts. These are the reasons which lead to the wrong diagnosis of disease. Like other diseases, the diagnosis of liver disease is a very complex task for human experts. Liver disease is rising day by day and is not discovered easily in its initial stage as the liver can work properly even when it is partially damaged. It is a very tough job for the physicians to recognize the disease from common symptoms [1]. Most of the symptoms are similar to other fever related ailments which result in the wrong diagnosis of disease. As other diseases dominate the liver disease due to which it is unable to identify.

Presently, research efforts are growing in the field of artificial intelligence to automatically diagnose the disease using medical knowledge, to provide the conscious prescription of further medical examinations and treatment. For the correct

* Corresponding author. Tel.: +91-998-877-9907.
E-mail address: amansingh.x@gmail.com.

diagnosis and classification of disease, various artificial intelligent techniques have been used. These include logistic regression, fuzzy logic, Artificial Neural Network (ANN), discriminant analysis, K-Nearest Neighbor (KNN), Radial Basis Field (RBF), Support Vector Machine (SVM), hopfield network and decision tree etc [2]**.** The result of these techniques depends on the type of dataset, their dimensions, etc. These techniques have poor explanation ability. Therefore, there is a requirement of other artificial intelligent methods which have good reasoning power.

This study accordingly provides a contribution to the process of liver disorder diagnosis by shortening the time through the use of other real-time algorithms or techniques. The two algorithms are involved in this study: CBR and GA. They remove the limitations of other intelligent techniques as both have good explanation ability and capacity to evaluate more complex cases. They put more weight on real-world cases instead of other domains. CBR acts as an advanced decision support system in clinical practice. CBR uses previous cases as an experience to solve and understand new problems. These CBR systems help doctors and physicians to check, analyze and repair their solutions. A doctor or physician gives input of complete description of particular domain situation and CBR system analyzes the whole cases with similar solutions and returns the output to the doctor [3]. The doctor can also modify the solution to fit for the current case and can retain them as a new solution in a case base. The fundamental reason for this framework is to serve as doctor diagnostic assistant. This framework helped in problem management and utilized as an instrument to help the youthful doctors to check their analysis.

In previous researches, the selection of appropriate instances and optimization of weights of features had been done independently. Optimization of these components simultaneously prompts to better execution as a contrast with single models. This study proposes the simultaneous optimization of these components using GA. GA deals with complex knowledge discovery problems. It is a stochastic search technique which searches for the best possible solution intelligently [4]. It depends on natural evolution and some kind of genetic operations. To endorse the effectiveness of an integrated model, we used three distinct medical cases related to liver disease and represent the outcomes shown by proposed model.

The remainder of this article is ordered as follows where Section 2 describes the prior studies of different models related to CBR and GA. Section 3 represents the detailed methodology of CBR. Optimization of feature weights and selection process of suitable instances simultaneously by using GA approach is also explained in section 3. Section 4 shows the experimental results, and the last section recapitulates the paper with brief conclusions.

2. Background and Related Work

To diagnose whether the person suffers from liver disease or not, number of artificial intelligent techniques are used like ANN, KNN, SVM, fuzzy logic etc [2]. These intelligent techniques perform a significant role in liver disease diagnosis. ANN is a powerful data-driven approach used to detect complex non-linear relationships between independent and independent variables. KNN is automatically non linear where linear and non-linear distribution data can easily be detected by KNN. It is also sensitive to outliers and removing them before using KNN.SVM shows high accuracy when problems are of high dimensions and are not linearly separable [5]. Fuzzy system is good in handling uncertainties and rules are produced to interpret relationships between input outputs. These techniques are criticized as they do not provide the good explanation for the output and cause overfitting. To remove these limitations CBR is applied and it is easy to apply and maintain. However, the prediction accuracy of CBR is very low as compared to other intelligent techniques. Therefore, in order to optimize the results a GA technique has been integrated with CBR and this hybrid GA-CBR model is used as a tool to improve the classification accuracy.

In this section, the prior studies showed the concept and applications of CBR and GA to assist clinicians in different disease diagnosis. Both strategies are used in a number of applications of a medical area to redesign the work of specialized doctors and to improve the quality and efficiency of therapeutic administrations.

2.1. CBR in medicine

CBR uses previous knowledge to solve the critical problems. A new problem can be resolved by finding the most comparable past issue and reuse the knowledge and information in that situation [3] . CBR research has been done in the medical domain. This methodology is widely used in diagnosis and prognosis of different diseases. S. Rodr ́guez developed CBR framework for classification and finding the leukemia patients automatically in light of microarray information which involves different algorithms that allow classification, filtering, and knowledge extraction [6]. A. Singh et al. reviewed different intelligent techniques and their applications including single CBR and GA techniques. Integration of CBR and GA with other techniques is also presented [2]. R. Lin developed an intelligent model to diagnose the liver disease by integrating CBR and Classification and Regression Tree (CART) methods [7]. V. Sch employed fuzzy matching techniques which are integrated with the CBR system for stress diagnosis and to handle uncertainty, ambiguity intrinsically exists in the

reasoning of medical practitioners [8]. D. Gu et al. proposed a mixed case retrieval technique to reuse the knowledge of dental medical records by using different algorithms as opposed to composing medicinal records physically and enhancing the quality and proficiency [9]. R. Lin et al. developed an intelligent model which includes ANN to predict the liver illness and shows an integration of Analytic Hierarchy Process (AHP) and CBR to find out liver sickness type [10]. C. Chuang compared five single models and to enhance the diagnostic performance, CBR was integrated with the other techniques where CBR- added integrated model represents best accuracy and sensitivity as compare to other models [11]. S. Petrovic et al. built up a CBR framework for producing dosage anticipates treatment of new tumor patients by catching the experience of oncologists in treating past patients [12].

V. E. Ekong et al. represented neuro-fuzzy-CBR driven decision support system in the identification of illness due to depression by utilizing the solutions of past cases [13]. D. A. Sharaf-el-deen et al. developed a new integrated approach by combining CBR with the Rule-Based Reasoning (RBR) approach and adaptation process are applied automatically by utilizing adaptation rules [14]. Z. Dong et al. introduced a medical decision support system based on CBR for identification of plausible tension type headache and to remove the confusion of overlapping between a probable migraine and plausible tension type headache [15]. P. Singh presented a knowledge support system based on CBR which improves the quality of life of asthmatic patients [16]. G. Khussainova et al. introduced a CBR system to provide radiotherapy treatment plan for patients suffer from brain cancer and to revise the retrieval mechanism by employing well-known clustering methods which have more success rate as compare to the original system [17]. R. M. Saraiva et al. presented hybrid CBR and RBR approach to support cancer diagnosis [18]. S. Banerjee et al. proposed decision support system based on CBR to classify the variations occur in pictures of the retina due to age related macular degeneration and diabetic retinopathy [19]. S. Chattopadhyay et al. introduced CBR based expert system for complex medical diagnosis which uses KNN algorithm to inquiry k closest comparative cases in light of the euclidean separation measure [20].

In general, All CBR methods have common four REs which involve Retrieve, Reuse, Revise, and Retain process [3,21]. Retrieve process means retrieve the best similar case from the historical case base library to answer the new issue. Reuse step is responsible for reusing the solution of a similar or most closely matched problem to answer the target problem. Revise step revise the solution using domain specific knowledge, if the solution is unsuccessful. After then retain the revised solution in the case base library which will be helpful for solving problems in future [3].

2.2. GA in medicine

GA is an effective approach for finding solutions to different problems using optimization techniques. It continuously tries to find various possible solutions using different genetic operators like selection, crossover, and mutation. In the first step, construct the initial population of individuals randomly, and the individuals are known as chromosomes in genetic space. The collection of genes represents a chromosome. The essential building block of a chromosome is gene and the position of a specific gene in the chromosome is characterized by locus. GA has one fitness function to evaluate the population in every generation. The different operators help to find possible solutions to problems. Selection operator is used to select individuals from the population for reproduction according to their fitness value [4,22]. Chromosomes with high fitness are utilized to remove low fitness chromosomes. However, selection does not create new individuals into the entire population alone. Crossover is a process in which a new pair of chromosomes is generated by exchanging part of genes between two chromosomes with high fitness values. The process and effectiveness of chromosome depend on the coding. The last operator is a mutation in which value of gene changes randomly.

Ghaheri et al. introduced applications of GA in disease diagnosis, prognosis, screening, treatment planning, health care management which helps physicians to analyze the applications of GA in their medical career [4]. S. Vinterbo et al. introduced a search technique to perform multi-disorder diagnosis which is based on GA [22]. Q. Yu et al. implemented a two stage integrated evolutionary approach which comprises genetic programming in the first stage and GA in the second stage on hepatitis and breast cancer datasets for better comprehension of undesirable medical events [23]. Y. Zhang et al. proposed a generic optimized technique to analyze the classification performance on liver disorders dataset [24]. C.Wu et al. characterized ultrasonic pictures of liver tissue into the typical liver, hepatoma, and cirrhosis by automatically selecting effective feature subset using GA- based feature selection method [25]. P. Pal et al. applied continuous GA methodology for developing intelligent computer-aided diagnosis system for heart disease diagnosis [26](Pal et al. 2013). A. G. Karegowda et al. presented an integrated model that combines GA and Back Propagation Network (BPN) to introduce and advance the connection weights of BPN to diagnose diabetes mellitus using PIMA Indian diabetes dataset [27]. E. Sreedevi et al. developed threshold GA for identification of diabetes disease using Minkowski distance method which is applied on PIMA Indian diabetes dataset [28]. J. M. D. Meth et al. introduced enhanced neuro-fuzzy system which is based on the GA in order to diagnose typhoid fever. The GA module works on the connection weights and supplied them to the hidden layer nodes to train the ANFIS- Adaptive Neuro-Fuzzy Inference system [29]. D. A. Antony et al. implemented GA to decrease

the dimensions of the original dataset which enhance the correctness and performance of classifiers namely J-48, Naïve bayes, and KNN, etc [30].

2.3. Integrated GA and CBR approach

H. Ahn et al. used CBR for customer classification and proved that optimization of feature weights' and selection of training examples for CBR using GA improve the classification accuracy at the same time as compare to naïve models [31]. H. Ahn et al. implemented hybrid CBR and GA approach for bankruptcy prediction [32]. P. Chang et al. developed a hybrid system by evolving GA and CBR for wholesaler's returning book prediction [33]. S. Kim et al. proposed hybrid CBR system with GA for determining the construction cost of building at preliminary design stage [21]. Y. Juan et al. used hybrid CBR and GA approach in housing customization to support decision-making [34]. Y. Park et al. proposed a new Cost Sensitive Case-Based Reasoning (CSCBR) method which can incorporate unequal misclassification costs into CBR and number of neighbors were enhanced by GA strategy [35].

3. Methodology

CBR is the most prominent information-driven approach among various machine learning techniques. It is easy to apply, provides the good explanation of each and every output and there is no overfitting. CBR method is applied on different real-world clinical datasets to analyze its prediction performance. However, in other applications CBR is always criticized because of its low prediction performance. An integrated GA-CBR model is proposed in this study to compare the effectiveness of this model with CBR in the prediction of liver disease. This model finds the impact of optimization algorithm on prediction results of CBR in liver disease. Optimization of weights of features and selection of appropriate instances for CBR is done simultaneously in this proposed model. The procedure involved in CBR and GA is explained as follows. The remaining part of this section describes the integrated GA-CBR model in prediction of liver illness.

3.1. Case-based reasoning

CBR is a problem-solving methodology of constructing knowledge based frameworks which are used to solve different problems. It is not an isolated technique which is used to solve its specific tasks. There are different techniques like nearest neighbor, fuzzy logic that follows the principles of CBR methodology. It adapts the prior solutions in order to find the solutions to target problems. In CBR terms, a case normally indicates a problem situation. A case represents features in matrix form. The basic principle behind the CBR methodology is that same type of solutions can be retrieved if the problem is of the same type [36]. CBR does not require in-depth analysis to solve the new target problems which make it differ from other artificial intelligent techniques. The reasoning procedure of CBR is explained in following steps:

Step 1: Indices and weights assignment

Input the training examples of real world liver disease dataset in medical record interface. In light of the expert conclusions from the gastroenterologist, a case index is worked to fuse different attributes. The weight is assigned to each feature to analyze the importance of feature which helps to measure the similarity [11]. If the weight of a feature is high, the significance of that feature is also considered to be high. Therefore, assignment of proper weights to the features is an important task.

Step 2: Case retrieval

In CBR, the purpose of this step is to retrieve the matching cases to the problems as fast as conceivable. Problem description is an input to the task of retrieval and output represents the cases that most closely match the target problem. Relevant cases are found according to the similarity in the features which can be measured by comparing the target with the other cases in the case base. Similarity calculation heavily depends on the weights allotted to case indices. Similarity calculation can be done for each quantitative or qualitative feature.

Qualitative similarity: The similarity between corresponding classes is defined by $sim(y,z)$ as described in Eq. (1). This is known as qualitative similarity and it mainly exists between categorical features.

$$sim(y,z)=\begin{cases}1\ if y=z, & similarity\ in\ the\ features\ belong\ to\ same\ class.\\ 0\ if y\neq z, & similarity\ in\ the\ features\ belong\ to\ different\ classes.\end{cases} \quad (1)$$

Quantitative similarity: It is a similarity measure of real or integer number with lower and upper bounds. The similarity measure value lies between 0 and 1.

$$if \quad s = t \; then \; similarity = 1$$
$$else \quad sim(s,t) = 1.0 - \left[\frac{|s-t|}{(up-lw)}\right] \tag{2}$$

where *s* is the query value, *t* represents the case value, *up* is the upper bound of number and *lw* is the lower bound of number.

Overall similarity: Similarity of both quantitative and qualitative features helps to calculate the overall similarity**.** The similarity score which is defined in Eq. (3) helps to retrieve the matching case from the case base library using the nearest neighbor computation.

$$similarity \; score = \frac{\sum_{j=1}^{k} W_j \, sim\left(f_j^I, f_j^R\right)}{\sum_{j=1}^{k} W_j} \tag{3}$$

where W_j is weight of index j and $\sum_{j=1}^{k} W_j = 1$. sim is similarity function for primitives. $f_j^I \; and \; f_j^R$ are values of input and retrieved cases for feature f_i. Retrieval task from the case base can be defined as the selection of small number of cases with the highest similarity to the query. The goal of matching task is to find the cases that are similar to the target problem and can find the solution to that problem. If this step has been executed completely then go to the step 3 else perform step 4.

Step 3: Case reuse

Reuse the case is the third step. After finding matching cases to the current case, the framework needs to reason as per the retrieved cases to locate a sensible and exact answer for the problem. Reusing the case can be done in two ways: the first one is considering diagnosis and classification application, where the solution of a retrieved case is directly used as the solution of input query [37]. The other way represents frameworks which depend on adaptation strategies. In this, the result of retrieved cases is not directly used as the solution to a new problem.

Step 4: Case adaptation

This step represents the tasks where we do not have possible solutions in the case base to solve a particular problem. Retrieve the matching cases, then find the similar solutions and use the difference between the input query and retrieved case in order to alter the retrieved solution for the target problem [38]. This is called the revising and optimization of the case. After performing this step go to step 5.

Step 5: Retain case

Reuse the cases and put the revised case in the liver disease case base for discovery of knowledge in future. This process is known as retention. Therefore, the solution of different problems can be used for further future problem solving mechanisms. However, sometimes this step leads to uncontrolled growth of the case base due to continuous retention of different cases which decreases the speed and ultimately reduces the performance of the system. Maintenance of the case base is necessary for these problems.

3.2. Genetic algorithm

Evolutionary Computation (EC) consist a number of techniques and approaches based on natural selection. GA is one of the algorithms that are generally applied to problems based on classifier systems and evolutionary strategies. It is an optimization technique which is inspired by natural genetics and natural selection. GA is an effective approach for finding solutions to different problems using optimization techniques. GA continuously tries to find various possible solutions using different genetic operators like selection, crossover, and mutation [4].CBR principally concentrates on the most proficient method to depict and retrieve cases. Weight is generated for each factor by using CBR and the best possible weight for each factor in CBR is discovered by GA. The steps for finding the best combination are as follows:

Step 1: Encoding in GA

The first step to solve a problem with GA is encoding of chromosomes. The process of encoding completely depends on the problem. Binary encoding is the most frequent method of encoding. Every chromosome is represented by a string of bits i.e. 0 or 1 and its combination is used in weight assignment to each factor [39]. This encoding is regularly not normal for some problems as some remedies must be required after crossover or mutation. Each factor influencing liver disease is assigned a weight with the combination of eight binary numbers.

Step 2: Initialize the population

The initial population is generated after the encoding step of GA. Every individual of population encodes a conceivable answer for a problem. The size of the population depends on the size of problems search space and computational time taken by it to evaluate each individual.

Step 3: Evaluation of chromosomes

After constructing the initial population, evaluation of each individual is done and a fitness value is assigned in accordance with the fitness function. The performance of every chromosome is evaluated by fitness function and it determines how closely it matches to the solution [39]. The comparison is made between the fitness values of a new chromosome and current chromosome. If fitness value for new chromosome is high then it will be reserved for new offspring.

In this, fitness value is calculated for each case of liver disease laboratory factors to predict whether the person suffers from liver disease or not. Fitness value for the training cases is as follows:

Objective function:

$$max\, M(A) = \sum_{j=1}^{n} R_j \tag{4}$$

s.t.

$$R_j = 1, if\, L_j = V_j$$
$$R_j = 0, if\, L_j \neq V_j$$

where $M(A)$ is an objective function to predict liver disease for a set of training cases A. L_j refers the predicted result of case j in training cases: If case j represents a person with liver disease then $L_j = 1$, otherwise the person is free from liver disease and $L_j = 0$. V_j refers the actual result of case j in training cases: If case j represents a person with liver disease then $V_j = 1$, otherwise the person is free from liver disease and $V_j = 0$. R_j defines the prediction and actual result comparison for case j, if the result is similar then $R_j = 1$else $R_j = 0$.

Calculation of L_j:

Let us consider,

Reference case set$C = \{c_1, c_2, \dots \dots \dots, c_m\}$, $p = 1,2,3, \dots, m$

Training case set $A = \{a_1, a_2, \dots \dots \dots, a_n\}$, $j = 1,2,3, \dots, n$

$$L_j = Z(c_p), if\ S_{jp} = \min_{p} [D(c_p, a_j)]$$

and

$$D(c_p, a_j) = \sqrt{\sum_{z=1}^{k} w_z \left(f_{pz} - f_{jz}\right)^2}, \qquad z = 1,2, \dots, k$$

where $k$is the total number of factors, S_{jp} is the degree of similarity between case j and case p of the training set A and reference set C. D represents the aggregation of separations between every weighted component of training and reference case set. $Z(c_p)$ refers the consequence of case p of reference set C which is most like to case j of training set A, c_p represents the diseased person then $Z(c_p) = 1$, otherwise $Z(c_p) = 0$. w_z is the weight of variable z in reference cases. f_{pz} represents the estimation of variable z of case p in reference cases. f_{jz} is the estimation of variable z of case j in training cases.

Step 4: Fitness value computation

The performance of every chromosome is evaluated by fitness function and it determines how closely it matches to the solution. The concept of fitness is related to the fact that highest the fitness value, highest it tends to propagate to the next generation(Chang et al. 2006). Fitness function finds the total number of accurate classifications for the entire dataset. The objective function for the problem of prediction of liver disease patients described in this research is to find the accuracy value. Thus, for a set of training cases an objective function acts as a fitness function.

$$fit(A) = M(A) = \sum_{j=1}^{n} R_j \tag{5}$$

Step 5: Reproduction/selection

Reproduction or selection is the first operator which is applied on the population. This operator uses the proportional selection of population. This step involves the roulette wheel selection method which means the probability of each

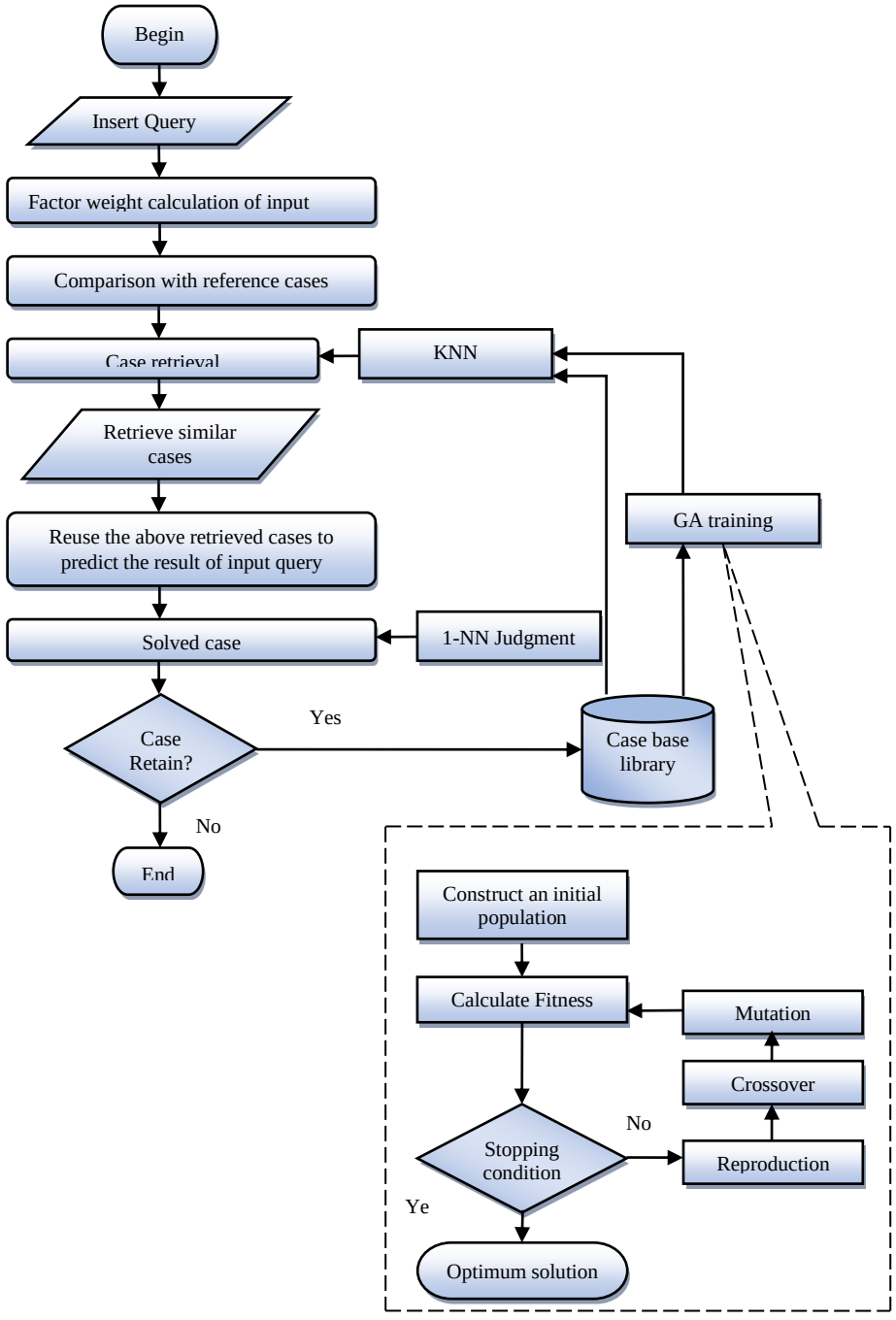

Figure 1. Integrated model combining CBR and GA

individual is equivalent to the fitness function $fit(y)$in every generation to the aggregate fitness value $\sum fit(y)$of whole population individuals [26]. This represents if the fitness value is higher, then the probability of being selected into next generation is also higher. Each chromosome y will be chosen to reproduce with probability $pr(y)$ is defined as:

$$pr(y) = \frac{fit(y)}{\sum fit(y)} \quad (6)$$

Step 6: Crossover

Crossover is the second operator applied on the population where two individuals are produced by selecting two parents. Crossover probability helps to perform two-point crossover operation in order to construct two new chromosomes.

Step 7: Mutation

The random substitution method is used in this study to perform mutation operation. It is the third operator which is applied to the population [39]. In this step, the chromosome which is to be mutated is replaced by new randomly generated chromosome with the low probability value.

Step 8: Update

The old individuals with poorer fitness value are replaced by new individuals after evaluating new individuals from crossover and mutation operations. This step is known as Replacement or Update step.

Step 9: Termination condition

The process of GA has repeated until the number of genetic cycles reached the maximum number of predefined genetic cycle number.

3.3. Integrated GA and CBR model

To develop an integrated framework of GA and CBR model there are some steps need to perform as demonstrated in Figure 1. In the first step, query is inserted into the case base related to liver disease dataset. The input query matches with the previous cases of liver disease dataset to find the proper solution to the problem [40,41]. Then determine the feature or factor weights of the input query. The assignments of factor weights in the proposed diagnosis model rely on the professional opinions of liver experts. The overall similarity heavily depends on the weights assigned to the case indices. After then compare the input query with reference cases from the historical case-base to find the matching case to given problem [21]. The comparison is made by similarity rule to predict whether the person suffers from liver disease or not.

In the case retrieval step, a set of similar cases from the previous cases can be retrieved by using KNN computation. KNN finds the k similar cases from the previous case base to find the solution to the new problem. For example, if the results of 5 similar cases are showing diseased persons then the result of the selected input query would be a diseased person for sure. In order to calculate the optimal weight of each factor, GA training is applied. The solution of retrieved case is reused directly to solve the problem of the input query. After then, put the revised case in the historical case-base for the discovery of knowledge in future and this procedure is known as retention. The termination condition occurs if we do not want to put the reused or revised case in the historical case base library.

4. Experiments and Results

4.1. Clinical data

In this study, we used three datasets related to liver disease where one dataset is Liver damage dataset which has been taken from AstraZeneca healthcare foundation and the other two are taken from University of California, Irvine (UCI) machine learning repository which includes BUPA liver disorder dataset, Hepatitis dataset. The liver damage dataset consists of five attributes ALP alkaline phosphate, ALT alanine aminotransferase, AST aspartate aminotransferase, TBL Total bilirubin, a dose which is a factor with levels A, B, C and D and 606 observations. The BUPA liver disorder dataset includes the data of 345 patients with 6 independent variables and 1 dependent variable. The independent variables are mcv, alkphos, sgpt, sgot, gammagt, drinks. The dependent variable is selector field with two categories – 1 or 2 on the basis of presence or absence of liver disease. The Hepatitis dataset has 155 instances with 19 independent variables and 1 dependent variable. The independent variables are age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology. On the other hand, the dependent variable represents a class with two categories 1 (DIE) or 2 (LIVE).

The name, representation and range values of Liver damage dataset, BUPA liver disorder dataset and Hepatitis dataset are described in Table 1, Table 2 and Table 3.

Table 1 Feature details of Liver damage dataset

Variable Name	Represented as	Meaning	Intervals
ALP	Integer	Alkaline phosphatase	15-129
ALT	Integer	Alanine aminotransferase	4-198
AST	Integer	Aspartate aminotransferase	5-104
TBL	Real	Total bilirubin at baseline	2.736-27.531

Table 2. Feature details' of BUPA liver disorder dataset

Variable Name	Represented as	Meaning	Intervals
Mcv	Integer	Mean corpuscular volume	65-103
Alkphos	Integer	Alkaline phosphotase	23-138
Sgpt	Integer	serum glutamic pyruvic transaminase	4-155
Sgot	Integer	Serum glutamic oxaloacetic transaminase	5-82
Gammagt	Integer	Gamma-glutamyl transpeptidase	5-297
Drinks	Real	Number of half-pint equivalents of alcoholic beverages drunk per day	0-20

Table 3. Feature details of Hepatitis dataset

Variable Name	Represented as	Meaning	Intervals
Age	Integer	Lifetime in years	7-78
Sex	Integer	Gender	Male, Female
Steroid	Integer	An organic compound with four rings	No, Yes
Antivirals	Integer	An agent that kills a virus	No, Yes
Fatigue	Integer	Tiredness	No, Yes
Malaise	Integer	Discomfort or illness	No, Yes
Anorexia	Integer	Loss of appetite	No, Yes
Liver Big	Integer	Enlarged or swollen liver	No, Yes
Liver Firm	Integer	Abnormalities of liver	No, Yes
Spleen Palpable	Integer	Enlargement of spleen	No, Yes
Spiders	Integer	Thin vessels form a web like shape	No, Yes
Ascites	Integer	abnormal accumulation fluid in the abdominal cavity	No, Yes
Varices	Integer	Enlarged veins	No, Yes
Bilirubin	Real	Yellow pigment in blood	0.3-8
Alk Phosphate	Integer	Protein found in body tissues	26-295
Sgot	Integer	Serum glutamic oxaloacetic transaminase	14-648
Albumin	Real	Protein found in blood	2.1-6.4
Protime	Integer	Prothrombin Time (PT) is a blood test	0-100
Histology	Integer	The study of the microscopic structure of tissues	No, Yes

4.2. Experimental results

In this study, classification is performed by case comparison. Leave-one-out is an evaluation method to measure the performance of an algorithm. It is a cross validation approach where training is performed on all data except for one point and prediction is made for that point [17]. The matching cases to the new case can be found by using K-NN method. K-NN computation represents k number of matching cases from the prior case-base. For both the simple and integrated model, K-

NN calculation is performed by changing the value of k from 1 to 9 by taking only odd numbers. After performing the experiment, the result predicts that 5-NN shows the best performance for the different cases. So we fix the value of k equal to 5 to find the 5 best matching cases to the new case from the reference cases. First, a CBR model is applied to all the three datasets to evaluate the accuracies and then simultaneous weight optimization is done by GA to enhance the performance evaluation of an integrated model. The metric accuracy of a model is calculated using following equation:

$$Accuracy = \frac{No.of\ correct\ predictions}{Total\ of\ all\ cases\ to\ be\ predicted}$$

$$= \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

For liver disease prediction, the values of TP, TN, FP and FN are interpreted as follows:
TP=It means True Positive i.e. CBR system decides the liver disease case, and domain expert decides a liver disease case.
TN=It means True Negative i.e. CBR system decides not a liver disease case, and also the expert decides not a liver disease case.
FP=It means False Positive i.e. CBR system decides a liver disease case, but the domain expert do not.
FN= It means False Negative i.e. CBR system decides not a liver disease case, but the domain expert decides it is liver disease case.

4.2.1. Performance evaluation of CBR model

CBR model is applied and the accuracies are measured for the above discussed three liver disease datasets. After loading the entire dataset, five most matching cases are found for every case with the help of K-Nearest Neighbor. Table 4 contains the details of a number of cases lie between the different accuracies range. For example, in liver damage dataset 3, 106, 482, 9 and 6 cases lie between (>=0 and <5), (>=10 and <15), (>=15 and <20), (>=20 and <25) and (>=25 and <30) accuracies range. In BUPA liver disorder dataset 2, 3, 45, 224, 68 and 3 cases lie between (>=0 and <5), (>=30 and <60), (>=60 and <65), (>=65 and <70), (>=70 and <75) and (>=75 and <85) accuracies range. In hepatitis dataset 46, 75 and 4 cases lie between (>=85 and <90), (>=90 and <95), (>=95 and <100) accuracies range and 30 cases are having accuracy 100%. Table 5 shows an average accuracy which is 17.32, 65.79 and 87.09 for Liver damage dataset, BUPA liver disorder dataset, and Hepatitis dataset.

Table 4. Case details of different datasets

Accuracies Range (%)	Number of liver damage dataset cases	Number of BUPA liver disorder dataset cases	Number of Hepatitis dataset cases
>=0 and <5	3	2	--
>=5 and <10	--	--	--
>=10 and <15	106	--	--
>=15 and <20	482	--	--
>=20 and <25	9	--	--
>=25 and <30	6	--	--
>=30 and <35	--	1	--
>=35 and <40	--	--	--
>=40 and <45	--	--	--
>=45 and <50	--	--	--
>=50 and <55	--	1	--
>=55 and <60	--	1	--
>=60 and <65	--	45	--
>=65 and <70	--	224	--
>=70 and <75	--	68	--
>=75 and <80	--	2	--
>=80 and <85	--	1	--
>=85 and <90	--	--	46

>=90 and <95	--	--	75
>=95 and <100	--	--	4
equal to 100	--	--	30
Total Cases	606 cases	345 cases	155 cases

Table 5. CBR Accuracy details of different datasets

Datasets	Liver damage dataset	BUPA liver disorder dataset	Hepatitis dataset
CBR Accuracy (%)	17.32	65.79	87.09

Table 6. GA-CBR Accuracy details of different datasets

Datasets	Liver damage dataset	BUPA liver disorder dataset	Hepatitis dataset
GA-CBR Accuracy (%)	24.42	68.98	94.19

4.2.2. Simultaneous weight optimization using GA (Integrated GA-CBR model)

CBR model has been widely used in the medical field. In this study, it has been used to diagnose the presence of liver disease or not. However, as a contrast with other machine learning methods CBR model is scrutinized in view of low prediction performance. Retrieval of cases from prior case base should be effective in order to obtain better results. Optimization of weights of features and appropriate instance selection simultaneously may lead to better performance than independent models. So here simultaneous optimization of weights and selection of appropriate instances for CBR is done by GA to improve the efficiency. The GA weight learning module is completely coordinated with the CBR.

The GA advances the weight structure. First, set the number of chromosomes to a small value i.e. 20 and set mutation rate to 0.05%. Every chromosome characterizes to the weights 1..N, as appropriate to a case base of cases each having N fields. The number of bits is 2 which represent 2*2=4 number of weight values. The weight 1 is represented by bit 00, weight 2 is represented by bit 01, weight 3 is represented by bit 10, and weight 4 is represented by bit 11. After setting all these parameters, run the CBR to evaluate each chromosome. The different genetic operators like crossover and mutation are applied to evaluate each new individual. Table 6 represents the accuracies of GA for different liver disease datasets which are giving better results as compare to CBR. The liver damage dataset, BUPA liver disorder dataset, and hepatitis dataset shows the accuracy of 24.42%, 68.98% and 94.19%. However, the accuracy result is best for hepatitis dataset i.e. 94.19%.

4.3. Observations and discussion

The motivation behind this work is the low performance shown by CBR method. GA based CBR model has been developed to enhance its performance. GA is a heuristic solution search procedure motivated by natural evolution. It is a powerful and adaptable approach that can be connected to an extensive variety of learning and enhancement issues. In this optimization of weights of features and selection of suitable instances have been done simultaneously [24]. This optimization is found to improve the retrieval results for complete CBR retrieval and its components. It is observed that optimization of these components at the same time leads to better results than separate optimization. As Table 5 shows the accuracy details of CBR whereas Table 6 shows the improved performance of CBR by an integrated GA-CBR model. This shows an effective impact of GA optimization in the prediction performance of CBR algorithm in liver disease. GA follows the concept of solution evolution by stochastically creating eras of solution population utilizing an objective function. GA is especially applicable to problems which are very large, discrete and non-linear in nature. They are applicable to features that add to the degree of complexity of solution. They do not break easily in the presence of noise or even if the inputs changed slightly. In addition to, searching can takes place in large state-space, multi-model state-space and n-dimensional state-space. These are some applicable reasons which make this model to optimize the performance of CBR.

Moreover, this model shows large variations in accuracies in terms of datasets. Table 6 shows the highest accuracy of 94.19% for hepatitis dataset as compare to other datasets. Here, features of hepatitis dataset add to the degree of complexity of solution. In addition to this, hepatitis dataset is discrete and non-linear in nature. Due to these reasons GA-CBR model is showing highest accuracy for hepatitis dataset as compare to other datasets. It is observed that the prediction performance of CBR has been optimized by introducing our proposed GA-CBR model. The applications of GA are purely available in hepatitis dataset and for the reason that this dataset shows highest accuracy.

5. Conclusion and Future Scope

Diagnosing disorders and diseases is one of the most difficult physician's responsibilities. An incorrect diagnosis can endanger a man's life and causes his death. In this regard, the use of different methods of artificial intelligence and expert system has become common and it is tried to minimize the error amount of these methods. In this article, CBR and GA techniques are analyzed for particular classification problem: The diagnosis of liver disease using clinical datasets. Two fundamental objectives are achieved: to begin with, the advancement of this specific application as an analysis device for human specialists; and second, an underlying correlation amongst GA and CBR, based on the outcomes obtained. These outcomes can give us a preparatory thought of the behavior of both procedures on similar classification problems.

In liver disease diagnosis, various machine learning techniques had been used because of their better performances. Nonetheless, some are reprimanded in view of its confinement's like- poor clarification capacity of the outcomes and overfitting. CBR model has been applied to overcome all those limitations, but it has weak performance. CBR model is showing an accuracy of 17.32%, 65.79% and 87.09% for Liver damage dataset, BUPA liver disorder dataset, and Hepatitis dataset. In this article, we have proposed an integrated GA and CBR framework that upgrades weights of elements and select appropriate cases at the same time. This hybrid model decreases the noise and removes the imprecise cases which cause incorrect prediction of performance. In addition, our model likewise finds suitable nearest cases for CBR methodology. This is done by using best possible weights of features in similarity calculation to enhance the prediction accuracy and performance of our model. An Integrated GA-CBR model is showing an accuracy of 24.42%, 68.98% and 94.19% for Liver damage dataset, BUPA liver disorder dataset, and Hepatitis dataset which are better as compare to single CBR model.

Of all the datasets, Hepatitis dataset has the highest accuracy which is 94.19% after applying the hybrid GA-CBR model. Hepatitis dataset has the highest accuracy because there are high linear correlations of features with class. Thus we can predict that the outcome of a model may depend on the experimental datasets. As our two datasets represent binary classification problem and one dataset represent multiclass classification problem. Moreover, it is concluded that weight optimized GA-CBR integrated model represents better results as compare to single CBR models. Our future work is centered on a few regions: the change in the representation of GA, the improvement on the distinctive phases of CBR.

References

1. Branch A, Azad I. "Using algorithms to predict liver disease Classification," *Electronics Information & Planning*, vol. 3, pp. 255-259, 2015.
2. Pandey B, Singh A. "Intelligent techniques and applications in liver disorders : A survey," *Int. J. of Biomedical Engineering and Technology*, vol. 16, no. 1, pp. 27-70, 2014.
3. Aamodt A. "Case-Based Reasoning : Foundational Issues , Methodological Variations , and System Approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994.
4. Ghaheri A, Shoar S, Naderan M, Hoseini SS. "The Applications of Genetic Algorithms in Medicine," *Oman Medical Journal*, vol. 30, no. 6, pp. 406–416, 2015.
5. Widodo A, Yang B-S. "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, 2007.
6. Rodríguez S. "Case-based reasoning as a decision support system for cancer diagnosis : A case study," *International Journal of Hybrid Intelligent Systems*, vol. 6, no. 2, pp. 97-110, 2009.
7. Lin R. "An intelligent model for liver disease diagnosis," *Artificial Intelligence in Medicine*, vol. 47, no. 1, pp. 53-62, 2009.
8. Sch V. "A Case-based Decision Support System for Individual Stress Diagnosis using Fuzzy Similarity Matching," *Computational Intelligence*, vol. 25, no. 3, pp. 180-195, 2009.
9. Gu D, Liang C, Li X. "Intelligent Technique for Knowledge Reuse of Dental Medical Records Based on Case-Based Reasoning," *Journal of Medical Systems*,vol. 34, no. 2, pp. 213–222, 2010.
10. Lin R, Chuang C. "A hybrid diagnosis model for determining the types of the liver disease," *Comput. Biol. Med.*, vol. 40, no. 7, pp. 665–670, 2010.
11. Chuang C. "Case-based reasoning support for liver disease diagnosis," *Artificial Inteligence in Medicine.*, vol. 53, no. 1, pp. 15–23, 2011.
12. Petrovic S, Mishra N, Sundar S. "A novel case based reasoning approach to radiotherapy planning," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 10759–10769, 2011.
13. Ekong VE, Ekong VE, Inyang UG, Onibere EA. "Intelligent Decision Support for Depression diagnosis based on Neuro-fuzzy CBR Hybrid Intelligent Decision Support System for Depression Diagnosis Based on Neuro-fuzzy-CBR Hybrid," *Modern Applied Science*, vol. 6, no. 7, pp. 79- 88, 2015.
14. Sharaf-el-deen DA, Ibrahim F. "A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems," *Journal of Medical Systems*, vol. 38, no. 9, pp. 1-11, 2014.
15. Yin Z, Dong Z, Lu X, Yu S, Chen X, Duan H. "A clinical decision support system for the diagnosis of probable migraine and

probable tension-type headache based on case-based reasoning," *The Journal of Headache and Pain*, vol. 16, no. 29, pp. 1-9, 2015.
16. Singh P. "ACS : Asthma Care Services with the Help of Case Base Reasoning Technique," in *Procedia Computer Science.*, vol. 48, pp. 561–567, 2015.
17. Khussainova G, Petrovic S, Jagannathan R. "Retrieval with Clustering in a Case-Based Reasoning System for Radiotherapy Treatment Planning," *Journal of Physics*, vol. 012013, pp. 1-11, 2015.
18. Saraiva RM, Bezerra J, Perkusich M, Almeida H, Siebra C. "A Hybrid Approach Using Case-Based Reasoning and Rule-Based Reasoning to Support Cancer Diagnosis : A Pilot Study," *Studies in health technology and informatics*, vol. 216, pp. 862–866, 2015.
19. Banerjee S, Roy A. "Case Based Reasoning in the Detection of Retinal Abnormalities using Decision Trees," *Procedia Comput. Sci.*, vol. 46, pp. 402–408, 2015.
20. Chattopadhyay S, Banerjee S, Rabhi FA, Acharya UR. " A Case- Based Reasoning System for Complex Medical Diagnosis," *Expert Systems*, vol. 30, no. 1, pp. 12–20, 2012.
21. Kim S, Shim JH. "Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry," *Canadian Journal of Civil Engineering*, vol. 41, no. 1, pp. 65–73, 2014.
22. Vinterbo S, Ohno-machado L. "A genetic algorithm approach to multi-disorder diagnosis," *Artificial Intelligence in Medicine*, vol. 18, no. 2, pp. 117–132, 2000.
23. Tan KC, Yu Q, Heng CM, Lee TH. "Evolutionary computing for knowledge discovery in medical diagnosis," *Artificial Intelligence in Medicine*, vol. 27, no. 2, pp. 129–154, 2003.
24. Zhang Y, Rockett PI. "A generic optimising feature extraction method using multiobjective genetic programming," *Applied Soft Computing*, vol. 11, no. 1, pp. 1087–1097, 2011.
25. Wu C, Lee W, Chen Y, Lai C, Hsieh K. "Expert Systems with Applications Ultrasonic liver tissue characterization by feature fusion," *Expert System Applications*, vol. 39, no. 10, pp. 9389–9397, 2012.
26. Pal P, Tomar S, Singh R. "Evolutionary Continuous Genetic Algorithm for Clinical Decision Support System," *African Journal of Computing & ICT*, vol. 6, no. 1, pp. 127–140, 2013.
27. Karegowda AG, Manjunath AS, Jayaram MA. "Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of PIMA Indian Diabetes," *Int. Journal of Soft Computing*, vol. 2, no. 2, pp. 15–23, 2011.
28. Sreedevi E, Padmavathamma PM. "A Threshold Genetic Algorithm for Diagnosis of Diabetes using Minkowski Distance Method," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 4, no. 7, pp. 5596–5601, 2015.
29. Meth JMD, Mg A, Ow S, Mo O, Awonusi O. "Enhanced Neuro-Fuzzy System Based on Genetic Algorithm for Medical Diagnosis," *Journal of Medical Diagnosis Methods*, vol. 5, no. 1, pp. 1–10, 2016.
30. Antony DA, Singh G. "Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis," *I.J. Intelligent Systems and Applications*, vol. 1, pp. 67–73, 2016.
31. Ahn H, Kim K, Han I. "Hybrid genetic algorithms and case-basedreasoning systems for customer classification," *Expert Systems*, vol. 23, no. 3, pp. 127–144, 2006.
32. Ahn H, Kim K. "Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach," *Applied Soft Computing*, vol. 9, no. 2, pp. 599–607, 2009.
33. Chang P, Lai C, Lai KR. "A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting," *Decision Support Systems*, vol. 42, no. 3, pp. 1715–1729, 2006.
34. Juan Y, Shih S, Perng Y. "Decision support for housing customization : A hybrid approach using case-based reasoning and genetic algorithm," *Expert Systems with Applications*, vol. 31, no. 1, pp. 83–93, 2006.
35. Park Y, Chun S, Kim B. "Artificial Intelligence in Medicine Cost-sensitive case-based reasoning using a genetic algorithm : Application to medical diagnosis," *Artif. Intell. Med.*, vol. 51, no. 2, pp. 133–145, 2011.
36. Watson I, "Case-based reasoning is a methodology not a technology," *Knowledge-Based Syst.*, vol. 12, no. 5–6, pp. 303–308, 1999.
37. Kaindl H, Śmiałek M, Nowakowski W. C."Case-based reuse with partial requirements specifications," in *Proceedings of the 2010 18th IEEE International Requirements Engineering Conference*, RE2010, 2010, pp. 399–400.
38. Mitra R, Basak J. "Methods of case adaptation: A survey," *International Journal of Intelligent Systems*, vol. 20, no. 6, pp. 627–645, 2005.
39. Kumar M, Husian M, Upreti N, Gupta D. "Genetic Algorithm: Review and Application," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 451–454, 2010.
40. Boral S, Chakraborty S, "A case-based reasoning approach for non-traditional machining processes selection," *Advances in Production Engineering & Management*, vol. 11, no. 4, pp. 311–323, 2016.
41. Khemani D, Selvamani RB, Dhar AR, Michael SM. "InfoFrax : CBR in Fused Cast Refractory Manufacture," *Proc. 6th Eur. Conf. CBR*, vol. 2416, pp. 275–283, 2002.

Sakshi Takkar received her B.Tech degree in Computer Science and Engineering from CT Group of Institutions, Punjab, India in 2015. Now she is pursuing her Masters of Technology in School of Computer Science and Engineering at Lovely Professional University, Punjab, India. Her areas of interest are Healthcare systems, Medical Experts Systems and Artificial Intelligence.

Aman Singh is working as an assistant professor in School of Computer Science and Engineering at Lovely Professional University, Punjab, India. He has about four years of teaching and research experience and his areas of interest are biomedical engineering, information security, cybercrime and computer forensics.