An Expert System based on SVM and Hybrid GA-SA Optimization for Disease Diagnosis

S. ANTO[1], Dr. S. CHANDRAMATHI[2], B. FEMINA[3]
[2]DEAN- Electrical Sciences
[1]Assistant Professor, [3]PG Scholar, Department of Computer Science and Engineering,
Sri Krishna College of Technology, Coimbatore, INDIA
[1]georgeantocse@gmail.com, [2]chandrasrajan@gmail.com, [3]feminayas@gmail.com

Abstract: - An accurate diagnosis of diseases by physicians is a tedious and challenging task. This challenge can be addressed by designing and implementing medical expert systems with utmost accuracy. This paper proposes a medical expert system based on Support Vector Machine (SVM) and hybrid Genetic Algorithm (GA) – Simulated Annealing (SA) for the diagnosis of a set of diseases by using the dataset of UCI machine learning repository. The SVM with Gaussian Radial Basis Function (RBF) kernel performs the classification process. The hybrid GA-SA is used for the selection of the most significant feature subset of the dataset and for the optimization of the kernel parameters of SVM. The performance of the expert system is analyzed using various parameters like classification accuracy, sensitivity and specificity. The proposed system is validated using different disease dataset like Pima Indian Diabetes (PID), breast cancer, hepatitis and cardiac arrhythmia. The classification accuracy of the proposed system is found to be superior to that of the other existing systems in the literature.

Key-Words: - Medical Expert System, Machine Learning, Genetic Algorithm, Simulated Annealing, Support Vector Machine.

1 Introduction

Machine learning focuses on the improvement of machine projects that develop and change when given new information. It is of two types: Supervised learning and Unsupervised learning. Supervised learning is the task of machine learning that infers a function from labeled training data. Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses. There are many examples of machine learning problems such as Optical Character Recognition (OCR), face detection, medical diagnosis and weather prediction. For decision making in data mining, classification is the most common technique used. They help users in understanding how the category relates to other categories.

Utilization of the classification system in medical diagnosis had drawn more attention. Assessment of information taken from patients and decisions of experts are the foundations of diagnosis. Determination of best features has more effect on the precision of the analysis framework in prediction. An automatic diagnosis problem can be approached via both single and hybrid machine learning methods. The proposed work diagnoses four diseases namely, diabetes, breast cancer, hepatitis and cardiac arrhythmia.

Diabetes mellitus is a condition that occurs when the body is not able to use glucose in general [1]. Glucose is the major source of energy for the body cells. The levels of glucose in the blood are restricted by a hormone called insulin. In diabetes, the pancreas does not make enough insulin or the body can't respond normally to the insulin. In the body of the diabetes affected patient, pancreas does not generate enough insulin or it does not respond normally to insulin. This increase the glucose levels in the blood, leading to symptoms such as increased urination and weight loss. Women have a tendency to be hardest hit by diabetes with a count of about 9.6 million as on date. About 8.8% of the aggregate women adult population of the 18 years of age and above in 2003 and this is almost a twofold increment from 1995 (4.7%). By 2050, the anticipated number of all persons with diabetes will have expanded from 17 million to 29 million [2].

Breast cancer is an unrestrained growth of breast cells. It is a collection of cancer cells that starts in the cells of the breast. A cancer can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign cancer is not considered cancerous: their cells are close to normal in form, they grow slowly, and they do not attack nearby tissues or spread to other parts of the body. Malignant cancer is cancerous. Left unchecked, malignant cells finally can spread beyond the original tumor to other parts of the body.

Diagnosing the tumors have turned into one of the trending issues in the medical field [3].

Viral hepatitis is another significant health issue around the globe [4]. Hepatitis is the aggravation and damage to hepatocytes in the liver and can be caused by autoimmunity, viruses, infections with fungi and bacteria, or exposure to toxins such as alcohol. Hepatitis diseases can be spread through blood transfusion and shared syringes.

The cardiac arrhythmia disease becomes one of fundamental diseases that undermine humans, particularly in developing countries such as USA and Canada. There are numerous approaches proposed to find heart arrhythmias by utilizing ECG signal. Electrocardiography is an essential device for recording an electrocardiogram (ECG) signals and the variability of bioelectric potential with respect to time as a human heart beats. The shape of the ECG waveform and heart rate determine the condition of cardiovascular health. It contains essential pointers to the nature of the ailment assaulting the heart [26].

2 Related Works

Varma et al [5] have proposed a computational intelligence approach for better diagnosis of diabetes. The modified Gini index-Gaussian fuzzy decision tree algorithm is proposed and is tested with Pima Indian Diabetes (PID) clinical dataset for accuracy. The obtained accuracy was 79.50%, which is low compared to other existing diabetes diagnosis systems.

Seera et al [6] have proposed a hybrid intelligent system for medical data classification. A hybrid intelligent system that consists of the Fuzzy Min-Max neural network, the classification and regression tree, and the random forest model is proposed. The system yields good generalization performance but is abysmally slow in test phase. The accuracy of the proposed system is 78.39%.

Calisir et al [7] have proposed an intelligent hepatitis diagnosis system using Principal Component Analysis (PCA) and Least Square Support Vector Machine (LSSVM).The simplest invariance could not be captured, unless the training data explicitly provides this information.

Kaya et al [8] have proposed a hybrid decision support system based on Rough Set (RS) and Extreme Learning Machine (ELM) for the diagnosis of hepatitis. Redundant features are removed from the dataset through RS approach and classification process is implemented through ELM by using the remaining features. The performance of the proposed system is decreased due to time delay.

Ruxandra Stoean et al [9] have proposed medical decision making model using SVMs, explained by rules of Evolutionary Algorithms (EA) with feature selection.SVMs successfully achieve high prediction accuracy due to a kernel-based engine and EAs can greatly accomplish a good explanation of how the diagnosis was reached. Similarly SVM concept is used for various datasets like diabetes, breast cancer and hepatitis as in literatures [21], [23], [24], [30], [31].

Zheng et al [10] have proposed breast cancer diagnosis based on feature extraction using a hybrid of k-means and SVM algorithms. The K-means algorithm is used to recognize the hidden patterns of the benign and malignant tumors separately. The membership of each tumor in these patterns is calculated and treated as a new feature in the training model.SVM is used to obtain the new classifier that differentiates the incoming tumors.

Ozcift et al [11] have proposed random forests ensemble classifier trained with data re- sampling strategy to improve cardiac arrhythmia diagnosis. A correlation based feature selection algorithm is used to select relevant features from cardiac arrhythmia dataset and RF machine learning algorithm is used to evaluate the performance of the selected features with and without simple random sampling to evaluate the efficiency of a proposed training strategy. The accuracy of the proposed system is 90.0%.There are few more works are found in literature which uses Neural Network (NN) classifier models for diabetes diagnoses like Artificial Neural Network (ANN) [20] and Linear Discriminant Analysis(LDA)- Artificial Neuro Fuzzy Inference Systems (ANFIS) [23]. For cardiac arrhythmia disease diagnosis from ECG signal data ANN [24], Multilayer Neural Network model (MNN) [26] is used. The breast cancer dataset solved using RST [30], Grid Algorithm [22] and CART [32] are also some of the works related to the proposed work.

This paper proposes a medical decision support system which uses SVM for classification and GA-SA for optimization as shown in fig 1.SVM is a kernel based statistical classification technique which is widely used to solve bi-class problems .In this work the SVM uses Gaussian kernel function. GA is an evolutionary algorithm which offers multi criterion optimization for higher dimensional pace problems. It is a popular local search method used for optimization. Simulated Annealing (SA) is a local heuristic approach which uses greedy technique to solve optimization problems. It is

based on the analogy between the simulation of the annealing of solids and the problem of solving large combinatorial optimization problems. In this work the hybridized GA-SA perform optimization to find both the most significant feature sub set and the kernel parameters of SVM.

3 Datasets Used

The proposed method includes four common diseases, namely diabetes, breast cancer, hepatitis and cardiac arrhythmia. All the data sets are available in the UCI machine learning repository.

The Pima Indian Diabetes (PID) dataset ,there are 268 instances in class '1' and 500 instances in class '0', where '1' means a positive test for diabetes and '0' is a negative test for diabetes. The total number of attributes is 8.

In the Wisconsin Breast Cancer Dataset (WBCD), there are 569 instances in which 357 cases are benign and 212 cases are malignant. Total number of attributes is 32.

In the hepatitis disease dataset, the presence or absence of hepatitis disease is predicted by using the results of various medical tests carried out on a patient. Hepatitis dataset contains 155 instances belonging to two different classes, 'die' with 32 instances and 'live' with 123 instances. The total number of attributes is 19.

In the arrhythmia dataset, there are 452 ECG records described by 279 attributes, out of which, 206 attributes are linear and the remaining 73 attributes are nominal. The first four attributes are age, sex, height and weight. The rest of the attributes are extracted from patients' ECG signals.

4 Proposed Scheme

The overall process of the proposed system is depicted in fig.1 which contains Data preprocessing, feature selection, classification and optimization.

4.1 Data Preprocessing

Data preprocessing is a data mining strategy that transforms raw data into a clear format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain numerous errors. By using mean algorithm, the incomplete data is filled.

4.2 Feature Selection

In feature selection, a subset of the original variables is discovered with mean to reduce and dispose the noise dimension. The principle behind feature selection is to pick a subset of input variables by taking out features with little or no predictive information. It transforms high-dimensional data into lower dimensions. When the input to an algorithm is too large to be processed and is suspected to be extremely redundant (much data, but not much information), then the input data is transformed into a reduced representation set of features.

Dimensionality reduction can be achieved by either taking out data that is nearly related to other data in the set, or joining data to make a small set of features [12]. In machine learning, the issue of supervised classification is concerned with utilizing labeled samples to impel a model that classifies objects into a limited set of known classes. The samples are described by a vector of numeric or nominal features. Some of these features may be unimportant or redundant. Avoiding unimportant or redundant features is essential because that they may have a negative impact on the accuracy of the classifier.

Selection of the most significant subset of features is an optimization problem. In this system, the feature selection is done using hybrid GA-SA local search mechanism which is explained in section 4.4.

4.3 Classification

Classification is a data mining function that assigns items in a collection to target classes. The objective of classification is to accurately predict the target class for each case in the data. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two sets: one for building the model (training) and the other for testing the model. SVM is a class of machine learning algorithms that can perform pattern

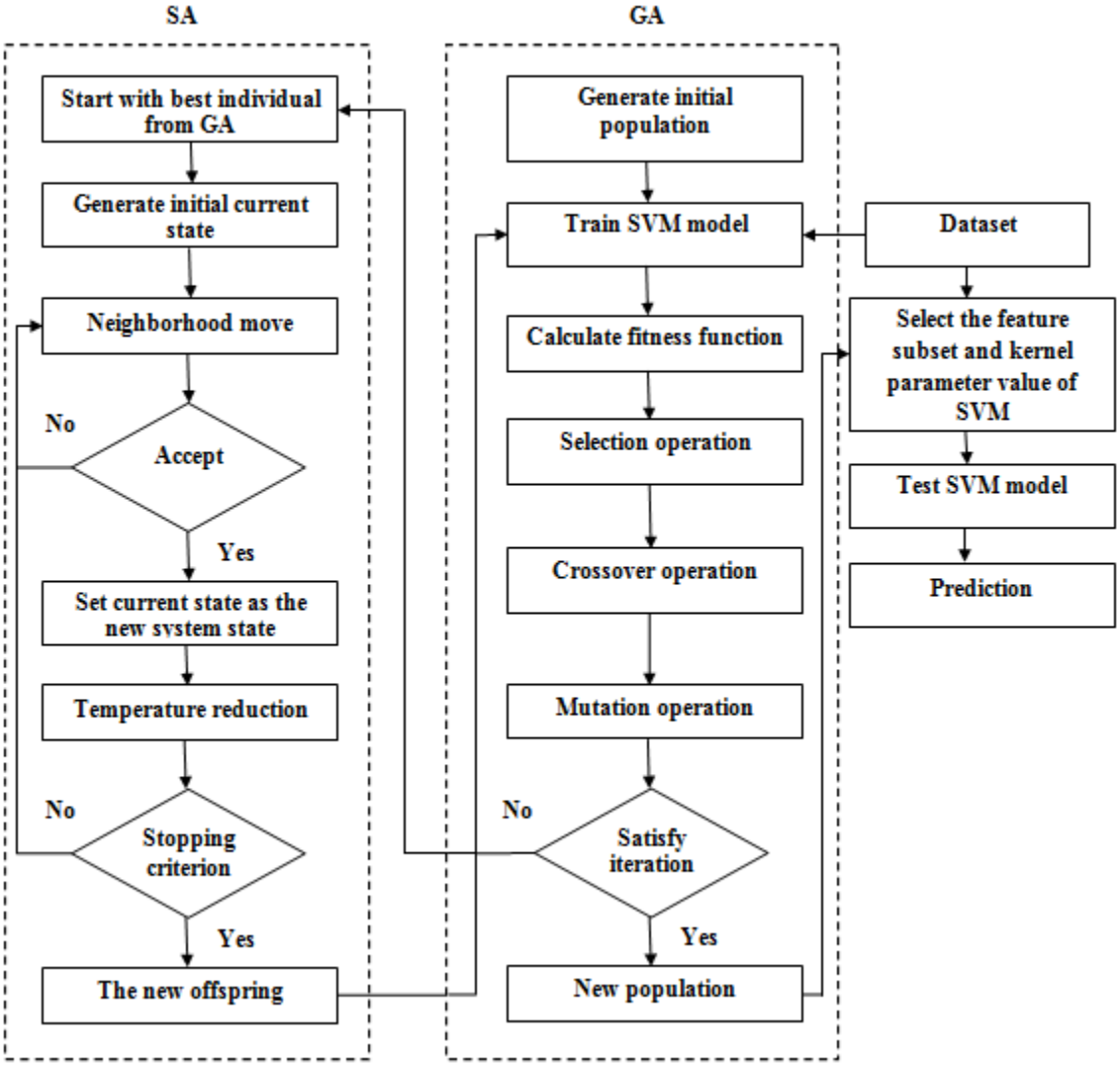

Fig.1 A flow diagram of the proposed GASA-SVM model

recognition and regression based on the theory of statistical learning and the principle of structural risk minimization [15].

Vladimir Vapnik invented SVM for searching a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin [16]. The margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples [17].

Given the training sample of instance-label pairs (x_i, y_i), $i = 1,\ldots, l$, $x_i \in R^n$, $y_i \in \{1,-1\}$, SVMs require the solution of the following (primal) problem [18]:

$$\min_{w,b,\varepsilon} \ \frac{1}{2} W^T W + C \sum_{i=1}^{l} \varepsilon_i$$

$$\text{Subject to} \quad y_i\left(W^T Z_i + b\right) \geq 1 - \varepsilon_i , \qquad (1)$$

$$\varepsilon_i \geq 0, i = 1, \ldots., l ,$$

where the training vector x_i is mapped onto a high dimension space by mapping function φ as $z_i = \varphi(x_i)$. $C > 0$ is the penalty parameter of the error term.

Usually, equation (1) is solved by solving the following dual problem:

$$\min_{\alpha} \ F(\alpha) \ \frac{1}{2} \alpha^T Q_\alpha - e^T \alpha$$

$$\text{Subject to} \ \ 0 \leq \alpha_i \leq C, \ \ i = 1, \ldots l \qquad (2)$$

$$Y^T \alpha = 0 ,$$

where 'e' is the vector of all 1's and 'Q' is an l by l positive semi definite matrix.

The $(i, j)^{th}$ element of Q is given by

$$Q_{i,j} \equiv y_i y_j K\left(x_i x_j\right),$$

The kernel function is given by equation3

$$K\left(x_i, x_j\right) \equiv \varphi^T(x_i)\varphi(x_i) \qquad (3)$$

$\{\alpha_i\}_{i=1}^{i}$ is Lagrange multipliers, and $W = \sum_{i=1}^{l} \alpha_i y_i \varphi$, (x_i) is the weight vector.

The classification decision function is,

$\text{sgn}(W^T\phi(x) + b) = \text{sgn}(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b)$ (4)

The kernel function $K(x_i, x_j)$ has manifold forms. In this paper, the Gaussian Kernel Function is used. Gaussian kernel function is expressed as follows:

$$K(x,x_i) = \exp(-\gamma\|x - x_i\|^2), \quad (5)$$

(or)

$$K(x, x_i) = \exp\left(-\frac{1}{\sigma^2}\|x - x_i\|^2\right) \quad (6)$$

Both the equations (5) and (6), which are in the same context, can transform parameter γ and parameter σ^2 . The Gaussian kernel parameter c is determined by $\gamma = \frac{1}{\sigma^2}$ .The parameters of SVMs with Gaussian RBF kernel refer to the pair : the error penalty parameter 'C' and the Gaussian kernel parameter γ ,usually depicted as (C, γ).

4.4 Optimization

The performance of a SVM classifier depends mainly on the values of the kernel function parameter, Gamma (γ) and penalty function parameter(C).Finding the best values of these two parameters, in order to achieve maximum classification accuracy of the SVM classifier, is a optimization problem. A hybrid GA-SA algorithm is used to solve this problem and to find the optimal values of C and Gamma.

4.4.1 Genetic Algorithm (GA)

GA is a model of machine learning which derives its behavior from a metaphor of some of the mechanisms of evolution in nature. The individuals represent candidate solutions to the optimization problem being solved. In GAs, the individuals are typically represented by n-bit binary vectors [13]. The resulting search space corresponds to an n-dimensional boolean space. It is assumed that the quality of each candidate solution can be evaluated using a fitness function. GAs use some form of fitness-dependent probabilistic selection of individuals from the current population to produce individuals for the next generation. The fitness function is used to search the best result. The selected individuals are submitted to the action of genetic operators to obtain new individuals that constitute the next generation.

Mutation and crossover are two of the most commonly used operators that are used with GAs to represent individuals as binary strings. Mutation works on a single string and generally changes a bit at random while crossover operates on two parent strings to deliver two off springs. Other genetic representations require the use of appropriate genetic operators. The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found.

- **Initial population:** Like in any step by step optimization drawback, the data of great starting parameter advantages the convergence speed of the algorithm. But such kind of data is rarely accessible. This leads to the coverage of large parts of the solution space due to the generation of the random initial population. Hence, to make the exploration of the solution space easier, a heterogeneous initial population is suitable.
- **Selection Step:** According to the fitness value, the individuals are positioned and most elevated rank is given to the best one. At that point, one solution is kept in the next generation.
- **Crossover Step:** The objective of this step is to assemble interesting features of a few solutions in new individuals by making combination of already held solutions.
- **Mutation Step:** This step brings the vital risk for effectively explore the solution space. Any purpose of this space can be reached, it is guaranteed. In addition, if a local optimum is obtained, then a too quick convergence to this neighborhood optimum will be kept away by mutation. Initially, it is set to a maximum value, then decreases to permit convergence and it expands again to avoid local optima.

Algorithm Steps:

Step 1: Randomly generate an initial source population of 'n' chromosomes.

Step 2: Calculate the fitness function f(x) of all chromosomes in the source population using equation (7).

$$\min f(x) = 100 * (x(1)^2 - x(2))^2 + (1 - x(1))^2 \quad (7)$$

Step 3: Create an empty successor population and then repeat the following steps until 'n' chromosomes have been created.

Step 4: Using fitness, select two chromosomes x_1 and x_2 from the source population.

Step 4.1: Apply crossover to x_1 and x_2 to obtain a child chromosome 'n'.

Step 4.2: By applying mutation to 'n', it deviant new offspring.

Step 4.3: Place new offspring in a new population.

Step 4.4: Replace the source population with the successor population.
Step 5: If the termination criterion is satisfied, stop the algorithm; otherwise go to Step 2.

4.4.2 Simulated Annealing (SA)

Annealing is a procedure in metallurgy, in which metals are gradually cooled to make them achieve low energy where they are very strong. In general, SA is a methodology where the temperature is reduced gradually, beginning from a random search at high temperature eventually getting pure greedy descent as it approaches zero temperature.

In optimization, the randomness should be appropriate to jump out of local minima and find regions that have a low heuristic worth. It is a probabilistic meta-heuristic for the global optimization issue of placing a good estimate to the global optimum of a given capacity in a large search space. It is regularly utilized when the search space is discrete [14].

Acceptance Function

The calculation chooses which answers to accept so we can better understand how it's able to avoid these local optimums.

Initially, the neighbor solution is checked to see whether it improves the current solution. If so, the current solution is accepted. Else, the neighbour solution is not better and hence the following couple of factors are considered.

- How greatly worse the neighbour solution is?
- How high the current temperature of our system is?

At high temperatures, the system is more likely to accept solutions that are worse.

Algorithm Steps:

Terminology:

X = Design Vector
fc = System Energy (i.e. Objective Function Value)
t = Temperature
Δ = Difference in system energy between two configuration vectors

Simulated Annealing Algorithm:

Step 1: Choose a random X_i, select the initial temperature t_1 and specify the cooling schedule.
Step 2: Evaluate $f_c(X_i)$ using a simulation model.
Step 3: Perturb X_i to get a neighboring Design Vector (X_{i+1}).
Step 4: Evaluate $f_c(X_{i+1})$ using a simulation model.
Step 5: If $f_c(X_{i+1}) < f_c(X_i)$, X_{i+1} is the new current solution.
Step 6: If $f_c(X_{i+1}) > f_c(X_i)$, then accept X_{i+1} as the new current solution with a probability using equation (8)

$$\exp(-\Delta/t) \text{ where } \Delta = f_c(X_{i+1}) - f_c(X_i) \quad (8)$$

Step 7: Reduce the system temperature according to the cooling schedule.
Step 8: Terminate the algorithm.

Neighborhood Search (Hybrid GA –SA)

In the hybrid SA algorithm, the best obtained solution in each GA generation is transferred to SA in order to improve the quality of solution through neighborhood search to produce a solution close to the current solution in the search space, by randomly choosing one gene in a chromosome, removing it from its original position and inserting it at another random position in the same chromosome. According to this criterion, even when the value of the next solution is worse, the solution can be accepted based on current temperature to avoid algorithm to stick in a local optimum. In the cooling phase the new temperature is determined by the decrement function t.

5 Experimentation and Results

5.1 Performance Metrics

To evaluate the prediction performance of SVM classifier, the classification accuracy, sensitivity and specificity are defined and computed.

Confusion Matrix: Confusion matrix shows classifications and predicted. A confusion matrix for a classification problem with two classes is of size 2×2, as given in Table 1.

Table 1 Confusion Matrix

Predicted	**Actual**	
	Positive	**Negative**
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

- TP represents an instance, which is actually positive and predicted by the model as positive.
- FN represents an instance, which is actually positive but predicted by the model as negative.
- TN represents an instance, which is actually negative and predicted by the model as negative.
- FP represents an instance, which is actually negative but predicted by the model as positive.

Accuracy, Sensitivity and Specificity:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \times 100\% \quad (9)$$

Sensitivity is the true positive rate and specificity is the true negative rate. They are defined as follows:-

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (10)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP+TN}} \times 100\% \quad (11)$$

5.2 Simulation Results

The system was implemented using MATLAB 7.In this paper, four diseases are diagnosed by using MATLAB software. The UCI machine learning repository datasets namely, Pima Indian Diabetes (PID) dataset, Wisconsin Diagnostic Breast Cancer (WDBC) dataset, Hepatitis disease dataset and Cardiac arrhythmia disease dataset are used to validate the proposed system with 80 % tuples for training and the remaining 20 % for testing.

In GA, an average of 30 generations is taken. The best fitness value found is 0.1481 and the mean fitness values for the four dataset are also calculated as 0.17 for PID, 0.19 for breast cancer, 0.18 for hepatitis and 0.18 for cardiac arrhythmia. The creation of generation and convergence of GA for all the four dataset- PID, breast cancer, hepatitis and cardiac arrhythmia is given in fig 2, fig 4, fig 6 and fig 8 respectively.

For SA, the initial temperature is set as 6.455 and the final temperature as 0.333. The temperature of SA is gradually reduced from initial value to the final in 50 cycles. The stopping criterion for the algorithm is given in equation (8).The GA receives the best chromosome with the help of SA. The working of SA using non-convex function for the four medical dataset- PID, breast cancer, hepatitis and cardiac arrhythmia is shown in fig 3, fig 5, fig 7 and fig 9 respectively.

The values of C and gamma achieved is C=400 and gamma = 0.0524.

In the PID dataset, the total number of attributes is 8, out of which 5 attributes are selected. The selected attributes list is shown in the Table 2.

Table 2 Selected features of Diabetes Dataset

Selected Attributes
Number of times pregnant
Plasma glucose concentration
Body mass index
Diabetes pedigree function
Age

In the WDBC dataset, the total number of attributes is 32, from which 12 attributes are selected. The selected attributes list is shown in the Table 3.

Table 3 Selected features of Breast Cancer Dataset

Selected Attributes
Texture
Perimeter
Area
Smoothness
Compactness
Concavity
Concave Points
Symmetry
Radius SE
Texture SE
Perimeter SE
Smoothness SE

In the hepatitis disease dataset, the total number of attributes is 19, out of which 10 attributes are selected. The selected attributes list is shown in the Table 4.

Table 4 Selected features of Hepatitis Dataset

Selected Attributes
Sex
Steroid
Antivirals
Fatigue
Anorexia
Liver Big
Spleen Palpable
Spiders
Ascites
Varices

In the arrhythmia dataset, there are 279 attributes. The first four attributes are namely age, sex, height, and weight, and the rest are attributes extracted from patients' ECG signals. The selected attributes list is shown in the Table 5.

Table 5 Selected features of Cardiac arrhythmia Dataset

Selected Attributes
QRS duration: Average of QRS duration
T interval: Average duration of T wave, linear
P interval: Average duration of P wave, linear

Heart rate: Number of heart beats, linear
Of channel DI: Q wave
Of channel DI: R' wave, small peak just after R
Number of intrinsic deflections, linear
Existence of diphasic derivation of T wave, nominal
Of channel AVR: Q wave
Of channel V1: R wave
Of channel V3: Q wave
Of channel V4: Existence of ragged P wave, nominal
Of channel V6: R wave
Of channel DI: Q wave, linear
Of channel AVF: Q wave

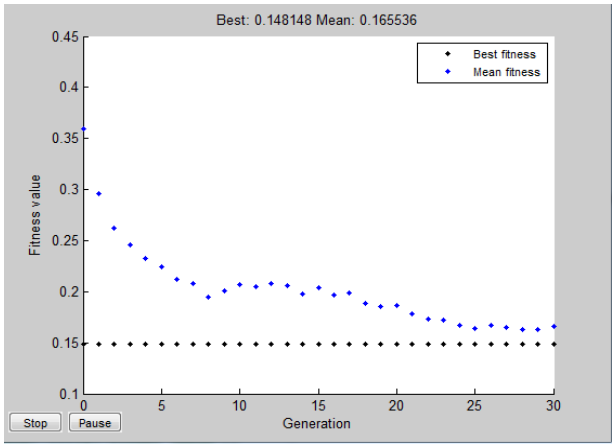

Fig.2 Generation creation using Genetic Algorithm for Diabetes dataset

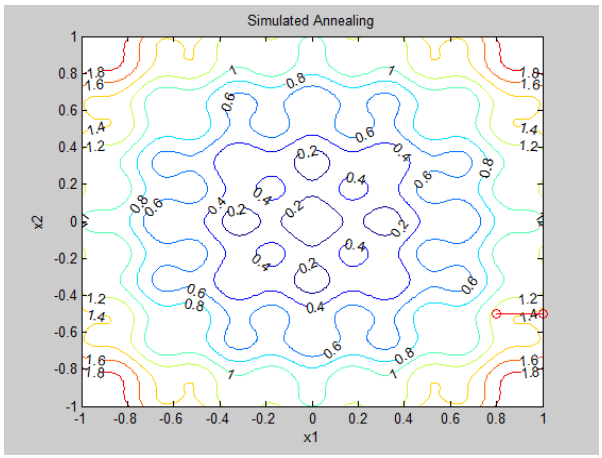

Fig.3 Neighborhood Search for Diabetes Dataset

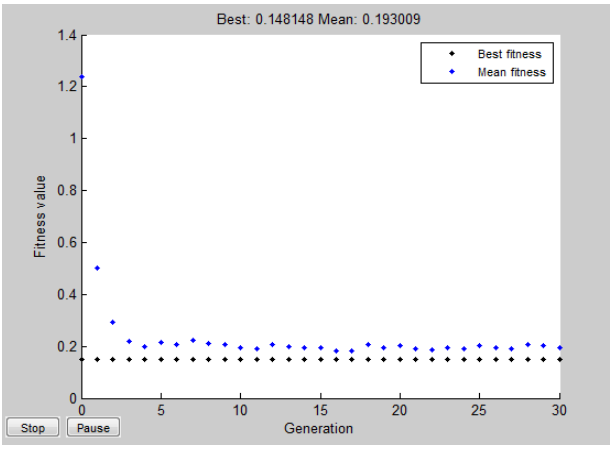

Fig.4 Generation creation using Genetic Algorithm for Breast Cancer Dataset

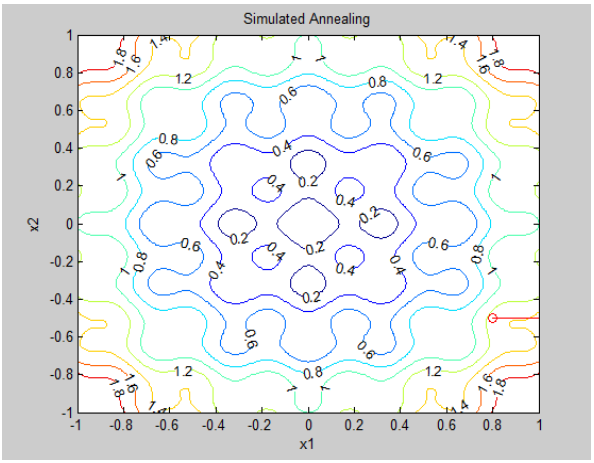

Fig.5 Neighborhood Search for Breast Cancer Dataset

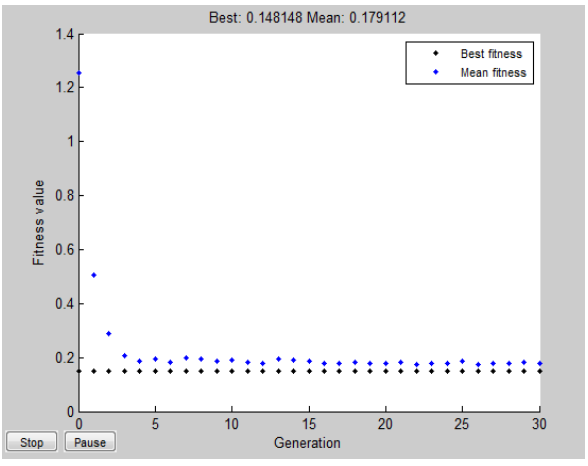

Fig.6 Generation creation using Genetic Algorithm for Hepatitis Dataset

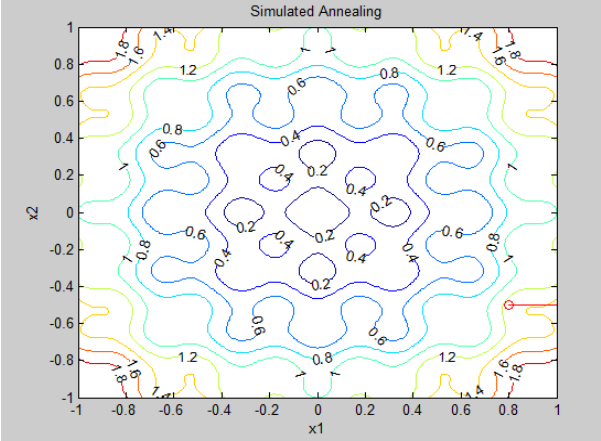

Fig.7 Neighborhood Search for Hepatitis Dataset

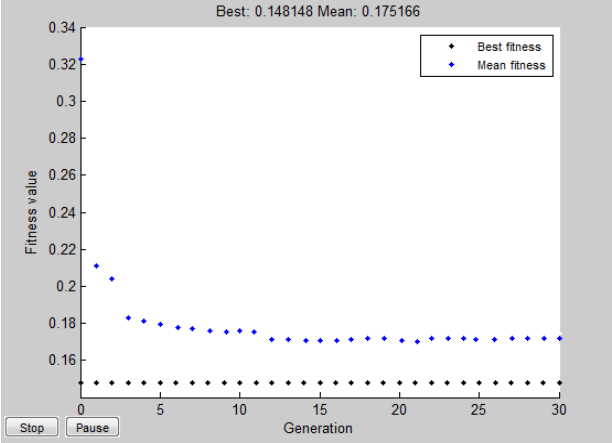

Fig.8 Generation creation using Genetic Algorithm for Cardiac Arrhythmia Dataset

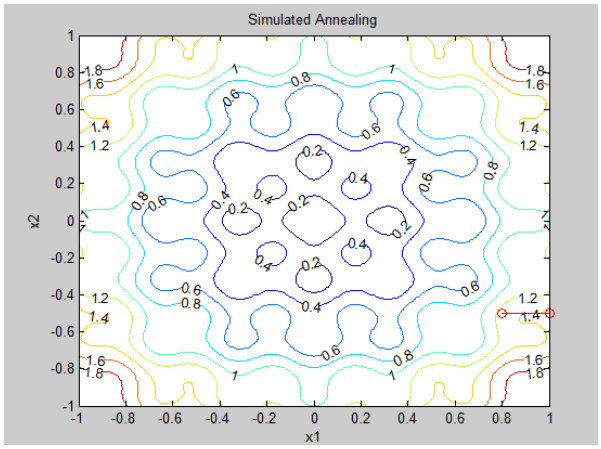

Fig.9 Neighborhood Search for Cardiac Arrhythmia Dataset

6 Results and Discussion

Using 80-20 train-test combination, the proposed system is constructed and tested. The hybrid GA-SA optimization module is used to perform feature selection. The SVM classifier kernel parameters are also optimized using the GA-SA algorithm. The optimized kernel parameters 'C' and gamma ensure the best performance of the classifier. The performance of the system is measured using the metrics like classification accuracy, sensitivity and specificity. The details of the performance metrics obtained for all the four medical dataset is listed in table 6. The comparison of the performance of the proposed system, in terms of the classification accuracy, with the existing systems are given in table 7, table 8, table 9 and table 10.

Table 6 Accuracy, Specificity and Sensitivity of the Proposed System

Dataset	**GASA-SVM**		
	Accuracy (%)	**Specificity (%)**	**Sensitivity (%)**
Diabetes	91.2	89.3	97.5
Breast Cancer	93.60	88.32	97.21
Hepatitis	87	82	93
Cardiac Arrhythmia	89.3	83.41	94.72

Table 7 Comparison of Accuracy between the Proposed Method and the Other Methods for PID

Diabetes	SVM NSGA-II [19]	ANN [20]	SVM [21]	GA-SVM [22]	LDA-ANFIS [23]	**Proposed GASA-SVM**
Accuracy%	86.13	73.4	94	82.98	84.61	**91.2**

Table 8 Comparison of Accuracy between the Proposed Method and the Other Methods for Breast Cancer

Breast Cancer	SVM [29]	GA-SVM [29]	RST [30]	Grid Algorithm [22]	CART [32]	**Proposed GASA-SVM**
Accuracy%	75.87	76.57	85	90.78	93.5	**93.60**

Table 9 Comparison of Accuracy between the Proposed Method and the Other Methods for Hepatitis

Hepatitis	GA-SVM [27]	SVM-SA [28]	SVM [28]	KNN [29]	C4.5 [31]	**Proposed GASA-SVM**
Accuracy%	86.12	88.08	74	75.0	83.60	**87.0**

Table 10 Comparison of Accuracy between the Proposed Method and the Other Methods for Cardiac Arrhythmia

Cardiac Arrhythmia	ANN [24]	CFS-LM [25]	RF [11]	MNN Model [26]	GFFNN [26]	MLP [26]	**Proposed GASA-SVM**
Accuracy%	86.67	87.71	90	82.22	82.35	86.67	**89.3**

7 Conclusion

An expert system based on GASA-SVM is proposed for the diagnosis of various diseases. The hybrid GA-SA is used for the selection of the most significant feature subset of the dataset and also this module optimizes the kernel parameters of the SVM. Achieving the optimal values of C and Gamma leads to the improved classification accuracy of SVM. The SVM uses the Gaussian RBF. The performance of the system is evaluated using various performance metrics like accuracy, sensitivity and specificity. The classification accuracy of the proposed system is compared with that of existing systems. In the future the system can be used to perform diagnosis on real life clinical data.

References:

[1] M. Pradhan, R. K. Sahu, Predict the onset of diabetes disease using Artificial Neural Network (ANN), *Int. J. Comput. Sci. Emerg .Technol .2*, 2011, pp.303–311.

[2] Boyle, James P., Amanda A. Honeycutt, KM Venkat Narayan, Thomas J. Hoerger, Linda S. Geiss, Hong Chen, and Theodore J. Thompson, Projection of Diabetes Burden Through 2050 Impact of changing demography and disease prevalence in the US, *Diabetes care 24*, No. 11, 2001, pp.1936-1940.

[3] R. Siegel, D. Naishadham, & A. Jemal, Cancer statistics, CA: *A Cancer Journal for Clinicians*, Vol.62, 2012, pp.10–29.

[4] J.Cohen, The scientific challenge of hepatitis C, *Science*, Vol.285, 1999, pp.26–30.

[5] Varma, Kamadi VSRP, Allam Appa Rao, T. Sita Maha Lakshmi, and PV Nageswara Rao, A computational intelligence approach for a better diagnosis of diabetic patients, *Computers & Electrical Engineering*, Vol.40, No. 5, 2014, pp.1758-1765.

[6] Seera, Manjeevan, and Chee Peng Lim, A hybrid intelligent system for medical data classification, *Expert Systems with Applications* 41, No. 5, 2014, pp.2239-2249.

[7] Cslisir, Duygu, and Esin Dogantekin, A new intelligent hepatitis diagnosis system: PCA–LSSVM, *Expert Systems with Applications* 38, No. 8, 2011, pp.10705-10708.

[8] Kaya, Yılmaz, and Murat Uyar, A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease, *Applied Soft Computing* 13, No. 8, 2013, pp.3429-3438.

[9] Stoean, Ruxandra, and Catalin Stoean, Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection, *Expert Systems with Applications* 40, No. 7, 2013, 2677-2686.

[10] Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, *Expert Systems with Applications* 41, No. 4, 2014, pp1476-1482.

[11] Ozcift, Akin, Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis, *Computers in Biology and Medicine* 41, No. 5, 2011, pp.265-271.

[12] L.H. Cheng, J.W. Chieh, A GA - Based Feature Selection and Parameters Optimization for Support Vector Machines, *Expert Systems with Applications*, Elsevier, Vol.31, 2006, pp.231–240.

[13] D.E. Goldberg, Genetic Algorithm in Search, Optimization, and Machine Learning, Addison-Wesley, Boston, 1989.

[14] Javad Salimi Sartakhti, Mohammad Hossein Zangooei, Kourosh Mozafari, Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA), *computer methods and programs in biomedicine*, 2012, pp.570-579.

[15] S. Idicula Thomas, A.J. Kulkarni, B.D. Kulkarni, V.K. Jayaraman, & P.V. Balaji, A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in escherichia coli, *Bioinformatics*, 22, 2006, pp.278–284.

[16] Vapnik V, Support-vector networks, *Machine Learning*, 20, 1995, pp.273–297.

[17] J. Platt, Sequential minimal optimization: A fast algorithm for training support vector machines. In Advances In Kernel Methods – Support Vector Learning, Cambridge, MA, USA: MIT Press, 1998, pp. 185–208.

[18] S.S. Keerthi, & C.J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, *Neural Computation*, 15, 2003, pp.1667–1689.

[19] Zangooei, Mohammad Hossein, Jafar Habibi, and Roohallah Alizadehsani, Disease Diagnosis with a hybrid method SVR using NSGA-II, *Neurocomputing*, 136,2014, pp.14-29.

[20] Pradhan, Manaswini, and Ranjit Kumar Sahu, Predict the onset of diabetes disease using Artificial Neural Network (ANN), *International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004)*, 2, No. 2 , 2011.

[21] Barakat, Nahla, P. Andrew Bradley, and H. Mohamed Nabil, Barakat, Intelligible support vector machines for diagnosis of diabetes mellitus, *Information Technology in Biomedicine, IEEE Transactions on* 14, No. 4, 2010, pp.1114-1120.

[22] Huang, Cheng-Lung, and Chieh-Jen Wang, A GA-based feature selection and parameters optimization for support vector machines, *Expert Systems with applications* 31, No. 2 , 2006, pp.231-240.

[23] Dogantekin, Esin, Akif Dogantekin, Derya Avci, and Levent Avci, An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS, *Digital Signal Processing* 20, No. 4, 2010, pp.1248-1255.

[24] Jadhav, M. Shivajirao, L. Sanjay, Nalbalwar, and A. Ashok, Ghatol, Artificial neural network based cardiac arrhythmia disease diagnosis, In *Process Automation, Control and Computing (PACC), 2011 International Conference on*, IEEE, 2011, pp. 1-6.

[25] Mitra, Malay, and R. K. Samanta, Cardiac arrhythmia classification using neural networks with selected features, *Procedia Technology* 10, 2013, pp.76-84.

[26] M Jadhav, Shivajirao, Sanjay L Nalbalwar, and Ashok A Ghatol, Artificial neural network models based cardiac arrhythmia disease diagnosis from ECG signal data, *International Journal of Computer Applications* 44, No. 15, 2012, pp. 8-13.

[27] K.C. Tan, E.J. Teoh, Q. Yua, K.C. Goh, A hybrid evolutionary algorithm for attribute selection in data mining, *Expert Systems with Applications*, Elsevier, 2009.

[28] S.S. Javad, H.Z. Mohammad, M. Kourosh, Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA), *Computer Methods and Programs in Biomedicine*.8, 2012, pp.570-579.

[29] A.F. Atiya, A. Al-Ani, A penalized likelihood based pattern classification algorithm, *Pattern Recogn, 42*, 2009, pp.2684–2694.
[30] Hassanien, Aboul Ella, and Jafar MH Ali, Feature extraction and rule classification algorithm of digital mammography based on rough set theory, In *Available at www. wseas. us/ e-library/ conferences/ digest2003/ papers*, pp. 463-104.
[31] L. Ozyilmaz, T. Yildirim, Artificial neural networks for diagnosis of hepatitis disease, *International joint conference on Neural Networks (IJCNN)*, Portland, OR, USA, July20–24, IEEE, 2003, vol.581, pp.586–589.
[32] B. Ster, Dobnikar, A Neural network in medical diagnosis: comparison with other methods. EANN'96, 1996, pp. 427–430.