Classification of Hepatitis C Virus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine

Omar S.Soliman, Eman Abo Elhamd

Abstract— Hepatitis C Virus is one of the most dangerous diseases all over the world. It affects millions of people every year and could takes man's life. Many classification algorithms have been applied for its diagnoses and treatment. This paper proposes a hybrid classification system for HCV diagnosis, using Modified Particle Swarm Optimization algorithm and Least Squares Support Vector Machine. Principle Component Analysis algorithm is employed to extract features vector. As LS-SVM algorithm is sensitive to the changes of values of its parameters, Modified-PSO Algorithm was used to search for the optimal values of LS-SVM parameters in less number of iterations. The proposed system is implemented and evaluated on the benchmark HCV data set from UCI repository of machine learning databases. It was compared with another classification system, which utilized PCA and LS-SVM. The experimental results showed the superiority of the proposed system that was able to obtain classification accuracy of 98.86% versus 96.12% of the other system.

Keywords— Hepatitis C Virus (HCV); Principle Component Analysis (PCA); Particle Swarm Optimization (PSO); Least Squares Support Vector Machine (LS-SVM).

1. INTRODUCTION

Hepatitis C Virus (HCV) is a leading cause of chronic liver disease, cirrhosis, and hepatocellular carcinoma, as well as the most common indication for liver transplantation in many countries. There are about 170 million infected persons all over the world, which is about 3% of the global population [30]. Hepatitis C virus is the most severe type of all hepatitis types. It assaults the liver, causes swelling and redness in it. HCV will be transmitted to the new born baby if the mother is tainted from this virus. Diabetes is the leading cause of renal failure in many populations in both developed and developing countries. According to the professionals, the rate of death from hepatitis will be three times more in the next twenty years [9], [13], [26]. Least Squares-Support Vector Machine (LS-SVM) classier is one particular sample of Support Vector Machine (SVM) [45]. LS-SVM is used for finding an optimal hyper-plane, which separates various classes. It obtains this optimal hyper-plane by using maximum Euclidean distance to the nearest point. LS-SVM is a parametric algorithm. It has high sensitivity to the changes in the values of its parameters. On the other side, Modified-Particle Swarm Optimization (Modified-PSO) is a modified version of Particle Swarm Optimization (PSO) which is a heuristic algorithm inspired from the nature social behavior of birds. The main strength of PSO is its fast convergence, compared with other global optimization algorithms [5].

HCV Patients often stop pursuing the pegylated interferon (peg-IFN) and ribavirin (RBV) treatment because of the high cost and associated adverse effects. Besides that current interferon and ribavirin (IFN/RBV) therapy is only effective in 50%-60% of patients [48], [20], [42]. Many classification algorithms were applied on HCV patients' records to help in diagnosis and treatment of this disease, trying to classify the patients or predict their future state. The aim of this paper is to develop a classification system which could help physicians in diagnosis and treatment of HCV disease. The proposed system combines LS-SVM classier with a modified version of PSO optimization technique. The rest of this paper is organized as follows; section 2 articulates the problem background and related work. The proposed system is introduced in section 3 and experimental results are presented in section 4. The last section is devoted to the conclusion and further research.

2. PROBLEM BACKGROUND AND RELATED WORK

HCV is a worldwide health problem in both industrial and developing countries and its incidence is rising [26]. Many classification algorithms have been applied on this area trying to help in diagnosis and treatment for HCV patients by classifying them or predict their future state. This section will introduce some of these works. In [1], A review which presents recent findings on noninvasive alternatives for the diagnosis of fibrosis and cirrhosis in patients co-infected with HIV and HCV. An automatic diagnosis system that integrates PCA and ANN for classification of HCV is proposed in [4]. ANN algorithm was introduced as an aided non-invasive grading evaluation of hepatic fibrosis by duplex

- *Omar S.Soliman, is Currently an assistant lecturer in operations Research and Decision Support Department, Cairo University, Egypt. E-mail: dr.omar.soliman@gmail.com*
- *Eman Abo ElHamd is currently pursuing master's degree program in operations Research and Decision Support Department, Cairo University, Egypt, E-mail: e.aboelhamd@fci-cu.edu.eg*

ultrasonography [51]. A comparison between Back-propagation and Naive Bayes Classifiers to diagnose hepatitis disease was introduced in [17]. A proposed system which consists of ANN and multiple logistic regression (MLR) analysis models based on clinical factors to predict a 6-year incidence of metabolic syndrome including the insulin resistance index that calculated by homeostasis model assessment [14]. A framework to establish a prediction system for the personalized treatment of chronic HCV in a logistic regression model was proposed in [29] while an algorithm based on decision trees to determine the outcome of patients with acute liver failure was proposed in [28]. Decision tree learning algorithm was proposed for pre-treatment prediction of anemia progression in HCV infection [19]. In [47], the usage of a modern information system using medical informatics technology was proposed to support in diagnosis and treatment of the problems related to the liver disease. A hybrid model using PSO and Case Based Reasoning (CBR) for hepatitis disease diagnose was proposed in [27]. A neural network-based method was presented for the diagnosis and classification of patients infected with HCV [2]. In [6], an intelligent HCV diagnosis system using PCA and LS-SVM classifier for hepatitis diagnosis was proposed. A machine learning model for HCV diagnosis that hybridizes SVM and simulated annealing (SA) was proposed in [37]. An intelligent system using multiple linear regressions (MLR) and SVM for developing quantitative structure activity relationship (QSAR) model for HCV was introduced in [35]. A machine learning methods to predict HCV non-structural proteins 5B polymerase inhibitors was proposed [50]. In [10], Markov model was been used to estimate the lifetime costs and quality adjusted life years in two treatment strategies (a standard duration therapy and truncated therapy). In [25], they proposed a specific HCV evolution and response to the combined interferon and ribavirin therapy based on Bayesian Networks (BN), Linear projection (LP) and Self-Organizing Tree (SOT) models. The serum Fourier Transform infrared spectroscopy for noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis C was applied in [38]. A prediction model for genetic polymorphism and viral factors for chronic hepatitis C was proposed in [23]. In [18], Data Mining techniques were applied to reveal complex interactions of the risk factors and clinical features profiling associated with the staging of Non-hepatitis B virus/non-hepatitis C virus related hepatocellular carcinoma, While Data Mining was used to build a model which allows physicians to identify patients requiring HCC surveillance and those who benefit from IFN therapy to prevent HCC [22]. Also, In [23], a decision tree model was proposed for predicting the probability of response to therapy with peg-interferon plus ribavirin. Where, Decision tree with CART classification algorithm was developed to forecast response to therapy with number chronic hepatitis C patients in [12]. Three classification algorithms (naive Bayes, support vector machine and the C4.5 decision tree) were applied which could assess the associations between chronic fatigue syndrome(CFS) using genetic factors such as single nucleotide polymorphisms (SNPs) [15].In [6], a method based on PCA and LS-SVM Classifier for expert hepatitis diagnosis system was introduced. In [8], a model that predicts Egyptian patients' response based on their clinical and biochemical data for Peg-IFN and RBV using ANN and DT was proposed. In [3] A brief insight review on non-invasive methods for predicting liver fibrosis in HCV with their pros and cons to make easier for a clinician to choose better marker to assess liver fibrosis in HCV infected patients. A mathematical model of cross-immunoreactivity showed that the level of HCV intra-host adaptation correlates with the rate of cross-immunoreactivity among HCV quasispecies [41]. All sample patients were classified in a model based on RFP into 2 classes with rapid (RP) and slow (SP) progression to fibrosis [24].

2.1 Principal Component Analysis

PCA is an attributes extraction/reduction technique which is considered as one of the most prevalent and useful statistical method for dimensionality reduction. This method transforms the original data into new dimensions [2]. The new features are formed by taking linear combinations of the original features of the form:

$$H_1 = b_1' = b_{11}K_1 + b_{12}K_2 + \cdots + b_{1m}K_m \quad (1)$$

$$H_2 = b_2' = b_{21}K_1 + b_{22}K_2 + \cdots + b_{2m}K_m \quad (2)$$

$$H_p = b_p' = b_{p1}K_1 + b_{p2}K_2 + \cdots + b_{pm}K_m \quad (3)$$

In matrix style, we can write H = B.K, where K are known as the loading parameters. The new axes are attuned such that they are orthogonal to one another with utmost expand of information.

$$\mathrm{var}(H_i) = b_i' \sum b_i, \ I = 1, 2 \ldots p \quad (4)$$

$$\mathrm{cov}(H_i, H_j) = b_i' \sum b_j \ , I = 1, 2 \ldots p \quad (5)$$

K_1 is the first principal component holding the prime variance.

As the direct computation of matrix B is not achievable. So, in feature transformation, the first step is to ascertain the covariance matrix U which can be expressed as:

$$U_{m\times n} = \frac{1}{m-1}\left[\sum_{i=1}^{m}(K_i - \overline{K})'.(K_i - \overline{K})\right] \quad (6)$$

where $\overline{K} = \left(\frac{1}{m}\right)\sum_{i=1}^{m} X_i$. The next step is to determine the eigen values for the covariance matrix U. Eventually, a linear transformation is defined by n eigen vectors match up to n eigen values from a m-dimensional space to n-dimensional space $(n < m)$. Principal axes are also referred to as eigen vectors $E_1, E_2, \dots, E_m$ correspond to eigenvalues$\lambda_1, \lambda_2, \dots, \lambda_n$. Generally, the first few principal components hold most of the information. Analysis of variances proportion represents the total number of principal components that should be retained from the dataset. Algorithm-1, illustrates the main steps in PCA technique.

Algorithm-1: PCA

```
Step1: Recover basis: Calculate $XX^T = \sum_{i=1}^{t} x_i {x_i}^T$ and let $U =$
        eigenvectors of $XX^T$ corresponding to the top d
        eignvalues.
Step2: Encode training data: $Y = U^T$where $Y$ is a $d \times t$ matrix
    of encoding of the original data.
Step3: Reconstruct training data: $X = UY = UU^T X$
Step4: Encode test example: $y = U^T x$ where $y$ is a $d-$
dimensional encoding of x.
Step5: Reconstruct test example: $\hat{x} = Uy = UU^T x$.
```

2.2 Modified Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an algorithm Inspired from the nature social behavior and dynamic movements and communications of insects, birds and fish [6], [23]. The main strength of PSO is its fast convergence, comparing with many global optimization algorithms like Genetic Algorithms (GA), Simulated Annealing (SA) and other global optimization algorithms. The key concept is dealing with changes in velocity. In general, the main idea of PSO is as follows. For the i^{th} particle in d dimension, it could update its velocity and position using (7), (8). Where r_1 and r_2 are two random numbers in the range [0, 1], V_{id} is the momentum, ω_{id} is the inertia weight, C_1 is the cognitive learning parameter and C_2 is the social collaboration parameter. $X_{id} = (x_{i1}, x_{i2}, \dots, x_{id})$ is the position of the i^{th} particle, $P_i = (p_{i1}, p, \dots, p_{id})$ represents the best previous position (i.e. the position with the highest fitness value).

$$V_{id} = \omega_{id} V_{id} + C_1 r_1 (p_{id} - X_{id}) + C_2 r_2 (p_{gd} - X_{id}) \qquad (7)$$

$$X_{id} = X_{id} + V_{id} \qquad (8)$$

Inertia Weight plays an important role in the process of providing balance between exploration and exploitation. It determines the contribution rate of a particles previous velocity to its velocity at the current time step. In [5] different types of inertia weights were mentioned like Constant, Random, Adaptive inertia weight and many other types. In [29] a modified version of PSO was proposed. The main idea of this modified version is as in the following equations. For the i^{th} particle in d dimention, it could update its velocity and position using (9) and (10)

$$V_{id} = \lambda[\omega_{id} V_{id} + C_1 r_1 (p_{id} - X_{id}) + C_2 r_2 (p_{gd} - X_{id})] \qquad (9)$$

$$X_{id} = X_{id} + (\omega V_{id}) \qquad (10)$$

Where λ is a convergence factor, which can be calculated using (11)

$$\lambda = \frac{2}{|2 - C - \sqrt{C^2 - 4C}\,|} \qquad (11)$$

Where $C = C_1 + C_2$

In the proposed Algorithm ω_{id} could be calculated using (12) where t is the iterator over all iterations and T_{max} is the maximum number of iterations. With the increasing of t, parameter ω will be decreased linearly from 0.9 to 0.4 [5].

$$\omega_{id} = 0.9 - \frac{t}{T_{max}} * 0.5 \qquad (12)$$

The Modified-PSO algorithm steps are illustrated in Algorithm-1 with random inertia weight [29].

Algorithm-1: Modified-PSO

```
Step 1: Initialize population of particles $X(t)$ which consists of
random positions $x_1, x_2, \dots, x_n$ and velocities $V(t)$ are made up
of the particle's initial velocity $v_1, v_2, \dots, v_n$ on n dimensions.
Step 2: Evaluate the fitness for each particle.
Step 3: For each particle, find the maximum fitness and
compare it to the best found so far (pbest), if $f(x_i) < f(pbest_k)$,
then $f(pbest_k) = x_i$
Step 4: Set $P_i$ equals to the location of the maximum fitness
value $X_i$
Step 5: Compare fitness evaluation with the population's
overall previous best. If current value is better thangbest, then
reset gbest to the current particle's array index and value.
Step 6: Calculate the convergence factor $\lambda$ using,(5)
Step 7: Calculate the Inertia weight $\omega_{id}$ using,(6)
Step 8: Update the position of the particle according to, (3)
and, (4) and the new population $X(t+1)$ will be generated.
Step 9: Adjust the acceleration of the particles using (13)
$$v_i = \begin{cases} V_{max} & \text{if } v_i > V_{max} \\ -V_{max} & \text{if } v_i < -V_{max} \end{cases} \qquad (13)$$
Step 10: Loop to step (2) until stopping criterion is satisfied
(Reach a maximum number of iterations $T_{max}$)
```

2.3 Least Squares Support Vector Machine

Least Squares-Support Vector Machine (LS-SVM) classifier is one particular sample of Support Vector Machine (SVM) [31], [24]. One could finds the solution in LS-SVM by solving a set of linear equations instead of a convex quadratic programming problem for classical SVMs, The main target of LS-SVM is finding an optimal hyper plane, which separates various classes. It obtains this optimal hyper-plane by using maximum Euclidean distance to the nearest point. The LS-SVM classifier maps the input vectors into a high dimensional feature space for non-separable data. Then, the LS-SVM classifier finds an optimal separating hyper-plane in this higher dimensional space [22].

Given a training dataset of N points $\{x_k, y_k\}^N{}_{k=1}$ with input data $x_k \epsilon R^n$ and output $y_k \epsilon R$, we consider the following optimization problem in primal weight space:

$$\min J(w,b)_{w,b,e} = \frac{1}{2}w^T w + \frac{1}{2}\gamma\sum_{k=1}^{N} e^2{}_k \qquad (14)$$

Such that

$$y_k - (w^T \varphi x_k + b) = e_k\,, k = 1,2,\dots N \qquad (15)$$

Where γ is a regularization factor, e_k the difference between the desired output y_k and the actual output, and $\varphi(.)$is a nonlinear function mapping the data points into a high dimensional Hilbert space; in addition, the dot product in the high-dimensional space is equivalent to a positive definite kernel function $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$. In primal weight space, a linear classifier in the new space takes the following form, where w the weight vector is and $b \,\epsilon\, R$ which called as the bias term.

$$y(x) = \text{sign}(w.\varphi(x) + b) \qquad (16)$$

The dual space of this primal space was found by solving the Lagrangian function in (17)

$$L(w,e,\propto) = J(w,e) - \sum_{k=1}^{N} \propto_k (w^T\varphi(x_k) + e_k - y_k) \qquad (17)$$

Where $\propto_k$ Lagrangian multipliers and are called Support Vectors. The optimal solution for objective function in, (17) must satisfy the following Karush-Kuhn Tucker (KKT) conditions [22].

$$\frac{\delta L}{\delta w} = 0 \rightarrow \ w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k) \qquad (18)$$

$$\frac{\delta L}{\delta w} = 0 \rightarrow \propto_k = \ \gamma e_k,\ k = 1, \dots, N$$

$$\frac{\delta l}{\delta w} = 0 \rightarrow \ w^T\varphi(x_k) + \ e_k - y_k = 0, k = 1, \dots, N$$

The linear system in (19) will results after elimination of w and e which generates the Support Vector $\propto'_k$

$$\left(K + \frac{I}{\sigma}\right)\alpha = y \qquad (19)$$

Where $y = [y_1, y_2, \dots, y_N]^T$, $\propto = [\propto, \propto_2, \dots, \propto_N]^T$ and $K \epsilon R^{NxN}$ is the kernel matrix. The resulting LS-SVM model for function estimation is as in (20), where $K(.,.)$ is the kernel function

$$y(x) = \sum_{k=1}^{N} \propto_k K(x, x_k) \qquad (20)$$

LS-SVM (Algorithm-2) was implemented using Radial Basis Function (RBF), (21) [22].

$$K(x, x_k) = \exp\left(-\frac{|x - x_k|^2}{\sigma^2}\right) \qquad (21)$$

Algorithm-2: LS-SVM:

Step 1: Load the training data set of n data points, $\{x_k, y_k\}^N{}_{k=1}$ where x_i is the i^{th} input vector and $y_i \epsilon R$ is the corresponding i^{th} target with values $\{-1, +1\}$.

Step 2: Generate random weights for each input data point.

Step 3: Determine the value of the bias term b and initialize the error e for each point randomly.

Step 4: Initialize γ and σ using random values.

Step 5: Search for values of e, w and b that minimize the objective function, (14) and, (15).

Step 6: Construct the Lagrangian function in, (17) with the solution that must satisfy the KKT conditions in the set of, (18).

Step 7: Calculate number of support vectors ($\propto$) using, (19).

Step 8: Training data for LS-SVM model could be classified using (20) with RBF kernel function, (21).

Step 9: Classify any new point by, (16) using RBF kernel function (21).

Step 10: Loop until stopping criteria is met, usually until reach the maximum number of iterations.

3. PROPOSED SYSTEM

The proposed system is composed of 4 main phases including Data Pre-Processing, Features Extraction, Parameters Optimization and Classification. These phases are described in Algorithm-4, where, its conceptual schema is demonstrated in Fig-1.

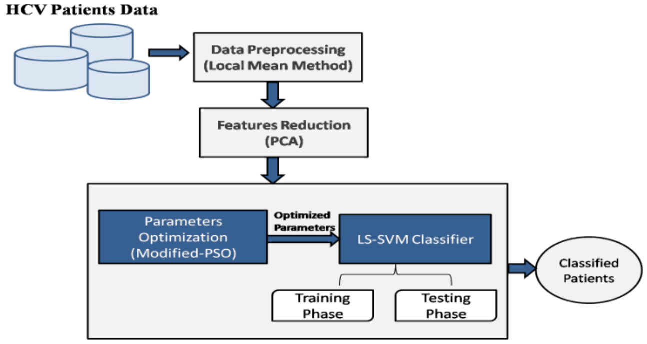

Fig. 1. Block diagram of the proposed system

- **Data Pre-processing**

Missing Data in the proposed algorithm are handled using Local Mean method. The main steps of Local Mean method is describes in Algorithm-5.

Algorithm-5: Local Mean Method:

Step 1: Determine the number of cells that we will calculate the average for (n).

Step 2: Given an empty cell that has an initial value (U = 0).

Step 3: Calculate the average of its n previous cells using equation (22).

$$U_t = \frac{U_{n-1} + U_{n-2} \dots U_0}{n} \qquad (22)$$

Step 4: Return number in the empty cell.

- **Features Extraction**

The proposed system is implemented and evaluated using the benchmark HCV data set from UCI repository of machine learning databases. The UCI dataset contains 154 HCV patients record and has 22 attributes as shown in Table-1. These attributes contains 12 binary and 10 attributes with discrete values; there are 32 cases out of 154 that die due to HCV.

Table 1: The set of attributes for UCI data set

Attribute No	Variable	Value
1	Fibroses Type.	F_0,F_1,F_2,F_3,F_4
2	AGE.	10,20,30,40,50,60,70,80
3	GENDER.	Male, Female
4	BILIRUBIN.	0.39, 0.80, 1.20, 2.0, 3.0, 4.0
5	ALK PHOSPHATE	33, 80, 120, 160, 200, 250
6	SGOT.	13, 100, 200, 300, 400, 500
7	ALBUMIN.	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
8	Weight.	67,75,85,90,100,110
9	PROTIME.	10, 20,... , 90
10	HISTOLOGY.	No, Yes
11	STEROID.	No, Yes
12	STEROID.	No, Yes
13	FATIQUE.	No, Yes
14	MALAISE.	No, Yes
15	ANOREXIA.	No, Yes
16	LIVER BIG.	No, Yes
17	LIVER FIRM.	No, Yes
18	PLEEN PALPABLE.	No, Yes
19	SPIDERS.	No, Yes
20	ASCITES.	No, Yes
21	VARICES.	No, Yes
22	CLASS.	Die, Live.

- **Parameters Optimization**

  The aim of this phase is to find the optimal values for the parameters of the LS-SVM classifier (The regularization factor (C) and Gaussian Kernel function (σ)) using Modified-PSO (Section 2.2).

- **Classification**

  Given the optimal values for the classifier's parameters, LS-SVM is utilized to classify the HCV patients into one of two classes (Live/Die).

4. EXPERIEMENTAL RESULTS

In the pre-processing phase, the Local Mean method (Algorithm-5) is used to clear the noise from HCV patient's data records and handle missing data. The PCA (Algorithm-1) is used to extract the most effective features in the diagnosis. For UCI dataset, the most effective and reduced features are 6 out of 22 as reported in Table-2. Modified-PSO algorithm was used to optimize two main parameters of LS-SVM (C which is the regularization factor and σ , the width of the Gaussian kernel).

Modified PSO algorithm was run on a data set of 152 records, about 152 random individuals in the search space is generated for 100 Iteration. The output of PSO is C= 200 and σ = 0.8. This set of optimized parameters are used as input to LS-SVM algorithm, seeking to find the optimal hyper-plan that separates the search space into two classes (Live/Die) by minimizing the optimization problem (18), (19). RBF kernel function was used in the classification process (21). In order to evaluate the performance of the proposed system, the classification accuracy was calculated using (23). Where TP and TN stand for True Positive and True Negative respectively, which are the proportion of positive and negative cases that were correctly identified respectively. Positive cases are the records with "Live" label and negative ones are with "Die" label. FP and FN stand for False Positive and False Negative which are the proportion of negative cases that were incorrectly classified as positive and the proportion of positive cases that were incorrectly classified as negative respectively [46], [1].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (23)$$

Table 2: Extracted features from PCA Algorithm for UCI dataset

Selected Variables	Variable Description
AGE	The smallest value of the age is 7 and the largest is 78
Bilirubrin	The yellow breakdown product of normal hemecatabolism. which is found in hemoglobin, a principal component of red blood cells.
ALK Phosphate	An enzyme which exists in the blood. It is responsible for removing phosphate groups from many types of molecules
SGOT	Its an enzeme which works as indicator for liver health.
Albumin	Any protein that is soluble in water. It also in moderately concentrated salt solutions.
PROTIME	Its predecessor of thrombin which is produced in the liver.
CLASS	Die/ Live.

The training phase was implemented using 10-fold Cross Validation (CV) method which breaks the data set into 10 sets of size n/10, train on 9 data sets and test on 1, then repeat this process 10 times and take a mean accuracy [31]. The average classification accuracy for LS-SVM is 98.86%, which obtained from the RBF kernel function (21). Table-3 shows the accuracy for every fold while applying 10-fold CV. The average accuracy over all 10 folds is 98.86%.

Table 3: Accuracy of 10-fold CV

Fold	Accuracy
1	96.9950%
2	95.9730%
3	98.8890%
4	99.9698%
5	98.9919%
6	98.9989%
7	99.9998%
8	99.9990%
9	99.9799%
10	98.8290%

In [6] the optimization of LS-SVM parameters was made by assuming a range for regularization factor (C) and Gaussian kernels (σ) such that $C \in [1,100000]$ and $\sigma \in [0.1, 25]$ and choosing the highest 20 of combinations of these values through 100 Iterations. The proposed system could avoid time consumed in these steps by optimizing LS-SVM parameters using modified-PSO. Table-4 demonstrates the superiority of the proposed system, which could obtain the same values for C and σ but with higher average classification accuracy and avoidance for doing large number of iterations while optimizing LS-SVM parameters.

Table 4: Comparison of the classification Accuracy

	regularization factor	Gaussian kernels	Accuracy
LS-SVM [6]	100	0.8	96.12%
Proposed System	100	0.8	98.86 %

Table 5: Classified Cases Vs Actual Cases

	Classified Values	
	Positive	**Negative**
Actual Values		
Positive:	142	10
Negative:	8	144

Table-5 demonstrates the confusion matrix of the proposed system. This allows visualization of the performance of an algorithm by illustrating how far the misclassified samples are from the actual classes and which degrees are better interpreted [44].

5. CONCLUSION AND FUTURE WORK

This paper introduced a classification system for HCV patients that integrates PCA, Modified-PSO and LS-SVM algorithms. The proposed system composed of 4 main phases including Data Pre-Processing, Features Extraction, Parameter Optimization and Classification. The PCA algorithm employed to extract the most effective HCV patients' features that support in diagnoses and treatment. The input parameters for LS-SVM were optimized using modified version of PSO algorithm. The LS-SVM algorithm was used to classify HCV patients into one of two classes (Live/Die). The proposed algorithm was implemented on benchmark HCV data set from UCI repository of machine learning databases. The proposed system was compared with another classification system, which utilized PCA with LS-SVM. The experimental results showed the superiority of the proposed system which could obtain classification accuracy of 98.86% while the other system obtained accuracy of 96.12%. As a future work, other optimization techniques could be used (i.e. Ant Colony System (ACS)). Also, other kernel functions could be applied in the classification phase.

REFERENCES

[1] Mohammed H Af, Abdel-Rahman Hedar, Taysir H Abdel Hamid, and Yousef B Mahdy. Ss-svm (3svm): A new classification method for hepatitis disease diagnosis. International Journal, 2013

[2] S Ansari I Sha J Ahmad and S Ismail Shah. “Neural network-based approach for the non-invasive diagnosis and classification of hepatotropic viral”. 2012

[3] Waqar Ahmad, Bushra Ijaz, Sana Gull, Sultan Asad, Saba Khaliq, Shah Jahan, Muhammad T Sarwar, Humera Kausar, Aleena Sumrin, Imran Shahid, et al. “A brief review on molecular, genetic and imaging techniques for hcv brosis evaluation”. Virol J, 8(1):53, 2011.

[4] JC Bansal, PK Singh, Mukesh Saraswat, Abhishek Verma, Shimpi Singh Jadon, and Ajith Abra-ham. "Inertia weight strategies in particle swarm optimization". In Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on, pages 633{640. IEEE, 2011

[5] James Blondin. Particle swarm optimization: A tutorial. from site: http://cs. armstrong.edu/saad/csci8100/pso tutorial.pdf, 2009.

[6] Duygu Calisir and Esin Dogantekin. A new intelligent hepatitis diagnosis system: Pca-lssvm. Expert Systems with Applications, 38(8):10705-10708, 2011.

[7] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011.

[8] Mahmoud ElHefnawi, Mahmoud Abdalla, Safaa Ahmed, Wafaa Elakel, Gamal Esmat, Maissa Elraziky, Shaima Khamis, and Marwa Hassan. Accurate prediction of response to interferon-based therapy in egyptian patients with chronic hepatitis c using machine-learning approaches. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pages 771{778. IEEE Computer Society, 2012.

[9] Centers for Disease Control and Prevention. Recommendations for prevention and control of hepatitis c virus (hcv) infection and hcv-related chronic disease.

[10] Ziad F Gellad, Andrew J Muir, John G McHutchison, William Sievert, Ala I Sharara, Kimberly A Brown, Robert Flisiak, Ira M Jacobson, David Kershenobich, Michael P Manns, et al. Cost-effectiveness of truncated therapy for hepatitis c based on rapid virologic response. Value in Health, 2012.

[11] Andrew M Gleason. Measures on the closed subspaces of a hilbert space. J. Math. Mech, 6(6):885-893, 1957.

[12] M Hassan, MI Abdalla, SR Ahmed, W Akil, G Esmat, S Khamis, and M ElHefnawi. The decision tree mode for prediction the response to the treatment in patients with chronic hepatitis c. New York Science Journal, 4(7), 2011.

[13] MAYO CLINIC HCV.
http://www.mayoclinic.com/health/hepatitis-c/ds00097, 1998-2013.

[14] Hiroshi Hirose, Tetsuro Takayama, Shigenari Hozawa, Toshifumi Hibi, and Ikuo Saito. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. Computers in Biology and Medicine, 41(11):1051 -1056, 2011.

[15] Lung-Cheng Huang, Sen-Yen Hsu, Eugene Lin, et al. A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data.J Transl Med, 7(1):81, 2009.

[16] Tahseen A Jilani, Huda Yasin, and Madiha Mohammad Yasin. Pca-ann for classification of hepatitis c patients.International Journal of Computer Applications, 14(7):1{6, 2011.

[17] Bekir Karlik. Hepatitis disease diagnosis using backpropagation and the naive bayes classifiers. IBU Journal of Science and Technology, 1(1), 2012.

[18] Takumi Kawaguchi, Tatsuyuki Kakuma, Hiroshi Yatsuhashi, Hiroshi Watanabe, Hideki Saitsu, Kazuhiko Nakao, Akinobu Taketomi, Satoshi Ohta, Akinari Tabaru, Kenji Takenaka, et al. Data mining reveals complex interactions of risk factors and clinical feature proling associated with the staging of non-hepatitis b virus/non-hepatitis c virus-related hepatocellular carcinoma. Hepatology Research, 41(6):564{571, 2011.

[19] Yoshihiro Kawamura, Shigeru Takasaki, and Masashi Mizokami. Using decision tree learning to predict the responsiveness of hepatitis c patients to drug treatment. FEBS open bio, 2:98-102,2012.

[20] Wan-Sheng Ke, Yuchi Hwang, and Eugene Lin. Pharmacogenomics of drug efficacy in the interferon treatment of chronic hepatitis c using classification algorithms. Advances and applications in bioinformatics and chemistry: AABC, 3:39, 2010.

[21] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23(1):89-109, 2001.

[22] Masayuki Kurosaki, Naoki Hiramatsu, Minoru Sakamoto, Yoshiyuki Suzuki, Manabu Iwasaki, Akihiro Tamori, Kentaro Matsuura, Sei Kakinuma, Fuminaka Sugauchi, Naoya Sakamoto, et al. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis c. Journal of hepatology, 56(3):602{608, 2012.

[23] Masayuki Kurosaki, Yasuhito Tanaka, Nao Nishida, Naoya Sakamoto, Nobuyuki Enomoto, Masao Honda, Masaya Sugiyama, Kentaro Matsuura, Fuminaka Sugauchi, Yasuhiro Asahina, et al. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis c using genetic polymorphism in il28b and viral factors.Journal of hepatology, 54(3):439{448, 2011.

[24] James Lara, Yury Khudyakov, F Xavier Lopez-Labrador, Fernando Gonzalez Candelas, and Ma-rina Berenguer. Hepatitis c virus genetic association to rate of liver brosis progression. In Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on, pages 1{1. IEEE, 2013

[25] James Lara, John E Tavis, Maureen J Donlin, William M Lee, He-Jun Yuan, Brian L Pearl-man, Gilberto Vaughan, Joseph C Forbi, Guo-liang Xia, and Yury E Khudyakov. Host-specific hcv evolution and response to the combined interferon and ribavirin therapy. InBioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, pages 102-109. IEEE, 2011.

[26] NIH National Institutes of Health MedilinePlus Trusted Health Infor-mation for You .. A service of the US. National Library of Medcine.
http://www.nlm.nih.gov/medlineplus/ency/article/000284.htm. 1998-2013.

[27] Neshat Mehdi, Sargolzaei Mehdi, Nadjaran Toosi Adel, and Masoumi Azra. Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. ISRN Artificial Intelligence, 2012, 2012.

[28] Nobuaki Nakayama, Makoto Oketani, Yoshihiro Kawamura, Mie Inao, Sumiko Nagoshi, Kenji Fu-jiwara, Hirohito Tsubouchi, and Satoshi Mochida. Algorithm to determine the outcome of patients

with acute liver failure: a data-mining analysis using decision trees.Journal of gastroenterology, 47(6):664{677, 2012.

[29] Hidenori Ochi, C Nelson Hayes, Hiromi Abe, Yasufumi Hayashida, Tomotaka Uchiyama, Naoyuki Kamatani, Yusuke Nakamura, and Kazuaki Chayama. Toward the establishment of a predic-tion system for the personalized treatment of chronic hepatitis c. Journal of Infectious Diseases, 205(2):204-210, 2012.

[30] International Journal of medical sciences. The natural history of hepatitis c virus (hcv) infection. (2):47-52, 2006.

[31] Rachel O'Reilly. Cross-validation for model selection in model-based clustering. 2012.

[32] Dakshata Panchal and Seema Shah. Article: An expert system for hepatitis b diagnosis using artificial neural networks. IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology (ICWET 2012), icwet(11):34-38, March 2012. Published by Foundation of Computer Science, New York, USA.

[33] Konstantinos E Parsopoulos and Michael N Vrahatis. Particle swarm optimization method for constrained optimization problems.Intelligent Technologies{Theory and Application: New Trends in Intelligent Technologies, 76:214{220, 2002.

[34] Vahdani Parviz, Aminzadeh Zohreh, Raoufy MohammadReza, Gharibzadeh Shahriar, Vahdani Golnaz, Fekri Sahba, Eftekhari Parivash, et al. Using artificial neural network to predict cirrhosis in patients with chronic hepatitis b infection with seven routine laboratory findings. Hepatitis Monthly, 2009(4, Autumn):271{275, 2011.

[35] Eslam Pourbasheer, Siavash Riahi, Mohammad Reza Ganjali, and Parviz Norouzi. Qsar study of c allosteric binding site of hcv ns5b polymerase inhibitors by support vector machine. Molecular diversity, 15(3):645{653, 2011.

[36] Salvador Resino, Matilde Sanchez-Conde, and Juan Berenguer. Coinfection by human immunodeciency virus and hepatitis c virus: noninvasive assessment and staging of fibrosis. Current Opinion in Infectious Diseases, 25(5):564{569, 2012.

[37] Javad Salimi Sartakhti, Mohammad Hossein Zangooei, and 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, 2011.

[38] Elodie Scaglia, Ganesh D Sockalingum, Juergen Schmitt, Cyril Gobinet, Nathalie Schneider, Michel Manfait, and Gerard Thie n. Noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis c using serum fourier transform infrared spectroscopy.Analytical and bioanalytical chemistry, 401(9):2919-2925, 2011.

[39] Xigao Shao, Kun Wu, and Bifeng Liao. Single directional smo algorithm for least squares support vector machines.Computational intelligence and neuroscience, 2013, 2013.

[40] Yuhui Shi and Russell Eberhart. A modified particle swarm optimizer. In Evolutionary Compu-tation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on, pages 69-73. IEEE, 1998.

[41] Pavel Skums, David S Campo, Zoya Dimitrova, Leonid Bunimovich, and Yury Khudyakov. Com-putational analysis and modelling of intra-host adaptation of hepatitis c virus: The role of immune cross-reactivity of hcv quasispecies. InComputational Advances in Bio and Medical Sciences (IC-CABS), 2013 IEEE 3rd International Conference on, pages 1-1. IEEE, 2013.

[42] Pavel Skums, David S Campo, Zoya Dimitrova, Gilberto Vaughan, Daryl T Lau, and Yuri Khudyakov. Modelling dierential interferon resistance of hcv quasispecies. In Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, pages 144-148. IEEE, 2011.

[43] Lindsay I Smith. A tutorial on principal components analysis. Cornell University, USA, 51:52, 2002.

[44] Catalin Stoean, Ruxandra Stoean, Monica Lupsor, Horia Stefanescu, and Radu Badea. Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis. Computers in Biology and Medicine, 41(4):238{246, 2011.

[45] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. Neural processing letters, 9(3):293-300, 1999.

[46] Pawe l Szewczyk and Miko laj Baszun. The learning system by the least squares support vector machine method and its application in medicine.Journal of Telecommunications and Information Technology, (3):109-113, 2011.

[47] Anna TSAKONA, Kallirroi PASCHALI, Dimitrios TSOLIS, and Georgios SKAPETIS. Hepatic liver diseases{methods for diagnosis and medical informatics for treatment support. Journal of Medical Informatics & Technologies, 17:211{218, 2011.

[48] Chun-Hsiang Wang, Yuchi Hwang, and Eugene Lin. Pharmacogenomics of chronic hepatitis c therapy with genome-wide association studies. Exper Pharmacol, 2:73{82, 2010.

[49] Weiwei Wang, Jie Cao, Hongke Lu, and Jian Wang. A default discrimination method for manu-facturing companies by improved pso-based ls svm.

[50] Lu Wei1 XUE Ying. Prediction of hepatitis c virus non-structural proteins 5b polymerase inhibitors using machine learning methods. Acta Physico-Chimica Sinica, 6:019, 2011.

[51] Li Zhang, Qiao-ying Li, Yun-you Duan, Guo-zhen Yan, Yi-lin Yang, and Rui-jing Yang. Artificial neural network aided non-invasive grading evaluation of hepatic fibrosis by duplex ultrasonography. BMC Medical Informatics and Decision Making, 12(1):55, 2012.

[52] M-H Zheng, K-Q Shi, X-F Lin, D-D Xiao, L-L Chen, W-Y Liu, Y-C Fan, and Y-P Chen. A model to predict 3-month mortality risk of acute-on-chronic hepatitis b liver failure using artificial neural network.Journal of viral hepatitis, 2012.