This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2690848, IEEE/ACM Transactions on Computational Biology and Bioinformatics

# Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients

Somaya Hashem[1], Gamal Esmat[2], Wafaa Elakel[2], Shahira Habashy[3], Safaa Abdel Raouf[1], Mohamed Elhefnawi[4], Mohamed El-Adawy[3], Mahmoud ElHefnawi[1, 5].

1) Informatics and Systems Department and biomedical informatics and chemo informatics group, Engineering Research Division and centre of excellence for advanced sciences, National Research Centre, Giza, Egypt.
2) Department of Endemic Medicine and Hepatology, Faculty of Medicine, Cairo University, Cairo, Egypt.
3) Communications, Electronics and Computers Department, Faculty of Engineering, Helwan University, Cairo, Egypt.
4) Communications and Computer Department, Faculty of Engineering, Modern University, Cairo, Egypt.
5) Center for Informatics, Nile University, Giza, Egypt.

## Abstract

***Background/Aim:*** Using machine learning approaches as non-invasive methods have been used recently as an alternative method in staging chronic liver diseases for avoiding the drawbacks of biopsy. This study aims to evaluate different machine learning techniques in prediction of advanced fibrosis by combining the serum bio-markers and clinical information to develop the classification models. ***Methods:*** A prospective cohort of 39,567 patients with chronic hepatitis C was divided into two sets – one categorized as mild to moderate fibrosis (F0-F2), and the other categorized as advanced fibrosis (F3-F4) according to METAVIR score. Decision tree, genetic algorithm, particle swarm optimization, and multi-linear regression models for advanced fibrosis risk prediction were developed. Receiver operating characteristic curve analysis was performed to evaluate the performance of the proposed models. ***Results:*** Age, platelet count, AST, and albumin were found to be statistically significant to advanced fibrosis. The machine learning algorithms under study were able to predict advanced fibrosis in patients with HCC with AUROC ranging between 0.73 and 0.76 and accuracy between 66.3% and 84.4%. ***Conclusions:*** Machine-learning approaches could be used as alternative methods in prediction of the risk of advanced liver fibrosis due to chronic hepatitis C.

**Index-terms:** Liver fibrosis prediction; Machine Learning Algorithm; Particle Swarm Optimization; Decision Learning Tree; Serum marker; Hepatitis C virus.

## 1. Introduction

In recent years, machine-learning techniques such as classification trees and artificial neural networks (ANN) have been used as prediction, classification, and diagnosis tools. Machine-learning techniques are used in the medical approaches to help using an invasive method in prediction and detection of diseases, such as prediction of fibrosis, cirrhosis, and prediction of response therapy in Hepatitis C Patients [1-3].

Hepatitis C is an infectious disease of the liver caused by the hepatitis C virus (HCV) [4] and is a major global cause of chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma. The prevalence of anti HCV antibody varies in different world countries with high reported rates in Egypt [5-6]. The assessment of liver fibrosis staging in Chronic Hepatitis C (CHC) is mandatory for the management of patients infected with the hepatitis C virus (HCV). It is essential to monitoring the prognosis of the disease, to establish the optimal timing for therapy, management strategies and to predict the response to treatment [7].

Liver biopsy was considered as a gold standard in staging liver fibrosis [8]. However, liver biopsy has potential risk due to its limitations including susceptible to sampling error and its invasive nature, addition to its highly cost for most of patients especially in periodic repeated for monitoring the diseases progress [6,9-10].

Therefore, in recent years the use of non-invasive methods as alternative in staging chronic liver diseases have significantly increased, in attempt to avoid the drawbacks of biopsy. According to Parkes et al [11], Serum markers of liver fibrosis offer an attractive alternative to liver biopsy. They are less invasive than biopsy, with no risk of complications, eliminate sampling and observer variability, easy to perform, and can be performed repeatedly [12]. Non-invasive methods in detection of fibrosis even based on indexes derived from serum markers [13-15], such as FIB-4 score and the aspartate aminotransferase (AST)-to-platelet ratio index (APRI) [16-17], or based on imaging techniques, such as using Transient Elastography (TE), which used ultrasound and vibratory waves for estimating the extent of liver fibrosis [18-20].

In this study, we will compare and evaluate the usefulness of different machine learning techniques in prediction of advanced fibrosis by combining the serum biomarkers and clinical information to develop the classification models. Particle swarm optimization, decision tree, multi-linear regression and genetic algorithm models for advanced fibrosis risk prediction were developed. The proposed models should be easy to perform, inexpensive, and give numerical and accurate results in real time. These models predicted the presence of advanced liver fibrosis with high accuracy and correlation coefficient especially with alternating decision tree and particle swarm optimization algorithms.

## 2. Patients and Methods

### 2.1 Patients

This study conducted in a Cohort of 39,567 chronic hepatitis C patients. The cohort data was enrolled in Egyptian National Committee for Control of Viral Hepatitis database in National Treatment Program of HCV patients in Egypt. They were 10741 female and 28826 male. The laboratory tests were performed at the same time of liver biopsy. By investigate and analyze the dataset of blood serum, it found to contain reported clinical information (such as age, gender and body mass index (BMI)), and contain histological findings (such as grade of fibrosis and the activity). In additional to laboratory tests( such as: albumin, total bilirubin, indirect bilirubin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alpha- fetoprotein (AFP), postprandial glucose test (PC%),international normalized ratio (INR), quantity of HCV_RNA, white blood cells (WBC) count, hemoglobin (Hb), platelet count, creatinine, serology finding, glucose, postprandial glucose test (PC%), and platelet count).

All data obtained on baseline, before starting antiviral therapy. Alcohol consumption was included in the questionnaire of the patients on baseline, most of the fields were missing or with denial of alcoholic consumption. Therefore and due to rarely consumption of alcohol by Egyptian people, a specific history of alcohol consumption was not considered as covariant. The study was done under informed consent that done by the national committee for control of viral hepatitis.

### 2.2 Liver Biopsy Histology

Liver histology is determined via METAVIR score [21] as assessed by local pathologists from Egypt. All patients underwent liver biopsy at baseline.

Total histological activity index and fibrosis scores (F0-F4) were recorded. According to the METAVIR system, fibrosis was staged on a scale from F0 to F4, as follows: F0: no fibrosis; F1: portal fibrosis, without septa; F2: few septa; F3: many septa without cirrhosis andF4: cirrhosis, respectively. F0, F1 and F2 were considered as mild to moderate fibrosis; whereas F3-F4 considered as advanced fibrosis [22].

### 2.3 Inclusion criteria and exclusion criteria

Inclusion criteria were: Age ≥ 18 years and ≤ 60 years. Positive HCV antibodies and detectable HCV RNA by PCR. Positive liver biopsy for chronic hepatitis with F1 METAVIR score and elevated liver enzymes or F2/F3 METAVIR score, naïve to treatment with PEG-IFN and RIB, hepatitis B surface antigen negativity, normal complete blood count, normal thyroid function, prothrombin concentration ≥ 60%, normal bilirubin, α-fetoprotein < 100 (ng/mL) and antinuclear antibody titer < 1/160.

Exclusion criteria were: Serious co-morbid conditions such as severe arterial hypertension, heart failure, significant coronary heart disease, poorly controlled diabetes (hemoglobin A1C > 8.5%), chronic obstructive pulmonary disease, major uncontrolled depressive illness. Solid transplant organ (renal, heart, or lung), untreated thyroid disease, history of previous anti-HCV therapy, body mass index (BMI) > 35 kg/m², known human immunodeficiency virus (HIV) co-infection, hypersensitivity to one of the two drugs (PEG-IFN, RIB), concomitant liver disease other than hepatitis C (chronic hepatitis B, autoimmune hepatitis, alcoholic liver disease, hemochromatosis, α-1 antitrypsin deficiency, Wilson's disease).

### 2.4 Statistical analysis, feature selection and classification

The data were statistically analyzed using MedCalc software and Microsoft Excel, while Matlab and Weka Softwares performed the PSO and DT learning algorithms. Data were reported as mean value ± standard deviation (SD). The relationship between variables and the presence of significant fibrosis has been assessed (P-value). The Kruskal-Wallis Test has been used for continuous variables with non-normal distribution. The Chi-square test has been used for categorical variables. Pearson correlation coefficients between fibrosis and each variable have been assessed.

We implemented several types of Machine learning techniques such as particle swarm optimization, genetic algorithm, multi-linear regression and decision tree learning algorithms. Decision tree algorithms; such as classification and regression tree (CART) [23], C4.5 [24], reduced error-pruning tree (REP), and alternating decision tree [25], have been build. We evaluated the performance of each of them on the datasets. The test set represents an external data set that was not used for training. The receiver operating curves (ROCs), sensitivities, specificities, predictive values and accuracies were applied to evaluate the performance of each model or technique on both the training and test sets.

### 2.5 Decision Tree Learning Algorithms

The Alternating decision tree (ADT) is a classification and predictive learning machine method. Traditional boosting decision tree algorithms such as CART [23] and C4.5 [24] have been successful in generating classifiers but at the cost of creating complicated decision-tree structures that are hard to interpret. Alternating decision tree (ADT) combines the simplicity of single decision tree with the effectiveness of boosting. It merges a number of weak hypotheses to induce a boosted one. At the same time, classifiers of this type are easy to interpret classification rules [25]. An alternating decision tree consists of decision nodes and prediction nodes. Decision nodes specify a collection of attributes. The branches between the nodes convey the possible values that these attributes can have in the observed samples. Prediction nodes have a numeric score, and they exist as both root and leaves. An instance is classified in an ADT by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed [26]. Figure 1(A) shows an ADT diagram; you can follow next steps on ADT tree for an instance:

1. Start with score = value of initial node.
2. Follow all paths evolved from the initial node, started with first decision node (ellipse shape) from the left.
3. For instance decision node, according to the condition on the arrow for instant attribute, go to the next prediction node (rectangle shape).
4. Update 'score' to be score= score+ instant prediction node value.
5. If the instant prediction node is not root node, Go to next decision node and repeat from step 3, or go to step 6.
6. If the instant prediction node is root node, go back to the nearest decision node (not followed before) and repeat from step 3 unless it is the final root node; you have to stop.
7. If the final score >= 0, then the patient classify to the positive class, and vice versa.

Proposed ADT model built from the algorithm is illustrated in figure 1(B).

### 2.6 Genetic Algorithms

Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover [27]. The algorithm starts with a set of individuals (called population) and advances towards the solution with three basic operations namely, selection, crossover and mutation. Each candidate solution has a set of properties (its chromosomes). Each chromosome consists of genes (e.g., bits), each gene being an instance of a particular allele (e.g., 0 or 1). The selection operator selects better individuals, allowing them to reproduce and pass on their genes to the next generation. The goodness of each individual depends on its fitness. The fitness function is always problem dependent. Crossover represents mating between individuals, it exchanges subparts of two chromosomes. Mutation randomly changes one or more gene values in a chromosome from its initial state, to maintain genetic diversity from one generation of a population of genetic algorithm chromosomes to the next. This generational process repeats until a termination condition was been reached, where either a maximum number of generations were been produced, or a satisfactory fitness level was been reached for the population. Figure 2 shows the flowchart of the genetic algorithm representing the order of its operations.

### 2.7 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population based stochastic optimization technique; developed by Dr. Eberhart and Dr. Kennedy in 1995[28]; inspired from the nature social behavior and dynamic movements and communications of insects, birds and fish. PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). It optimizes a problem by iteratively trying to improve a candidate solution by updating generations. However, PSO has fast convergence comparing with GA. PSO algorithm is initialized with having a population of random solutions (called particles) and moving these particles around in the search-space (d dimension) and searching for optima by updating generations. Particle could update its velocity ($V_{id}$) and position ($X_{id}$) using equations 1, 2. $X_{id} = (X_{i1}, X_{i2}, \dots , X_{id})$ is the position of the i$^{th}$ particle, $P_i = ( p_{i1},p_{i2}, \dots , p_{id})$ represents the best previous position, which has the highest fitness value. We can build the modified-PSO algorithm according to the following steps:

Step 1: Initialize a swarm population of N particles X(t), which consists of random positions and velocities in the search-space, where 't' is the iterator over all iterations.

Step 2: Initialize the particle's best-known position to its initial position.

Step 2: Evaluate the fitness for each particle.

Step 3: Update (pbest ) according to the maximum fitness of all particles.

Step 4: Set $P_i$ equals to the position of the maximum fitness value $X_i$.

Step 5: Update gbest according to a comparison of fitness evaluation with the population's overall previous best.

Step 6: Calculate the convergence factor λ using equation (3).

Step 7: Calculate the Inertia weight $\omega_{id}$ using equation (4). By increasing of t, Inertia weight w will be decreased linearly from 0.9 to 0.4.

Step 8: Update the position of each particle according to, (1) to generate the new population X(t+1).

Step 9: Adjust the acceleration of the particles according to equation (2).

Step 10: Go Back to step (2) until reach a maximum number of iterations $T_{max}$.

$$X_{id} = X_{id} + wV_{id} \tag{1}$$

$$V_{id} = \lambda[w_{id}V_{id} + C_1r_1(p_{id} - X_{id}) + C_2r_2(p_{gd} - X_{id})] \tag{2}$$

Where $r_1$ and $r_2$ are two random numbers in the range [0, 1], $V_{id}$ is the momentum, $w_{id}$ is the inertia weight, $C_1$ is the cognitive learning parameter and $C_2$ is the Social collaboration parameter. λ is a convergence factor, which can be calculated as:

$$\lambda = \frac{2}{\left|2-C-\sqrt{C^2-4C}\right|} \tag{3}$$

Where $C = C_1 + C_2$,

$$w_{id} = 0.9 - \frac{t}{T_{max}} * 0.5 \tag{4}$$

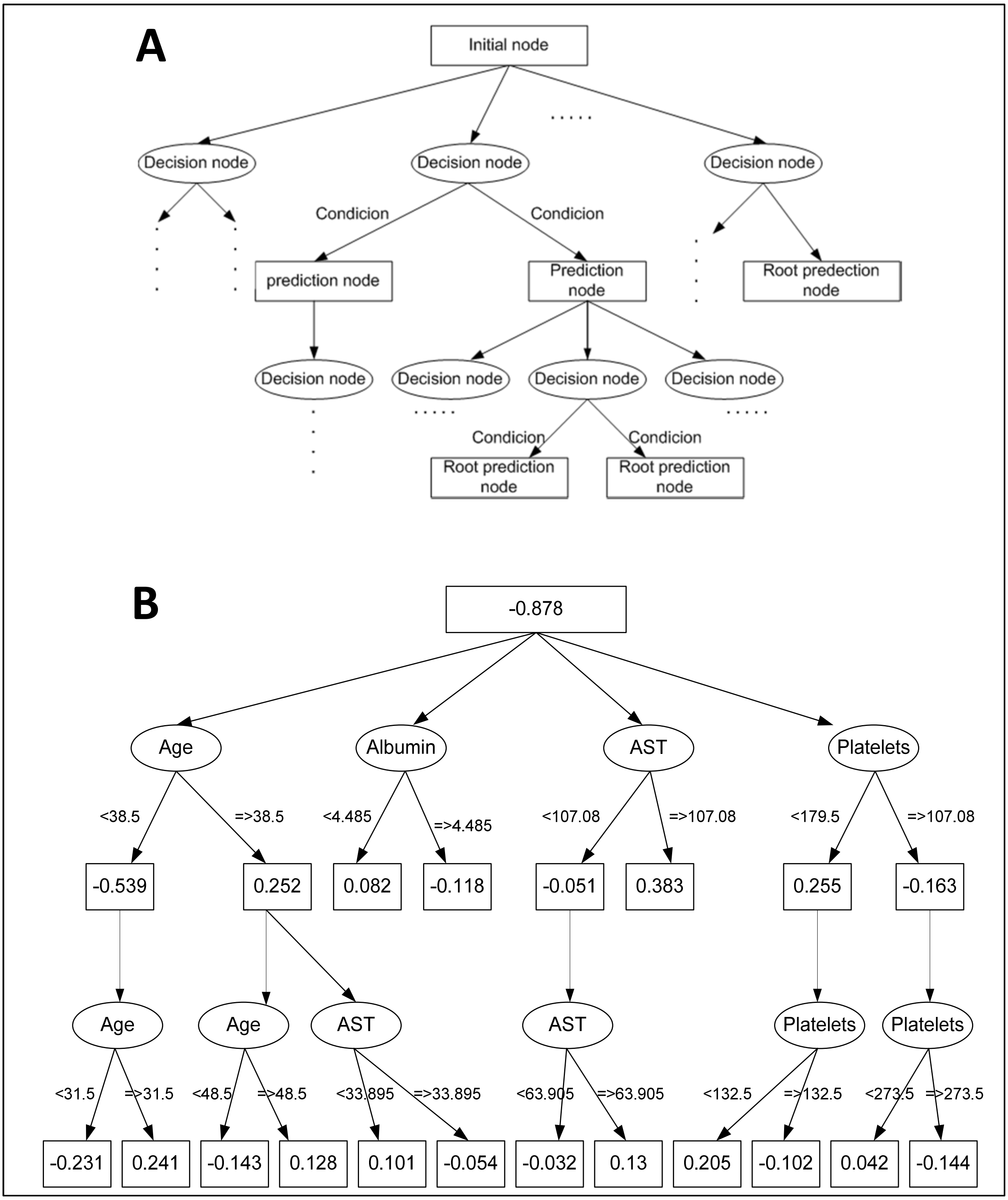

**Figure 1 Alternating decision tree.(A) Alternating decision tree diagram, An instance is classified by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed . (B) Proposed Model Alternating decision tree diagram.**

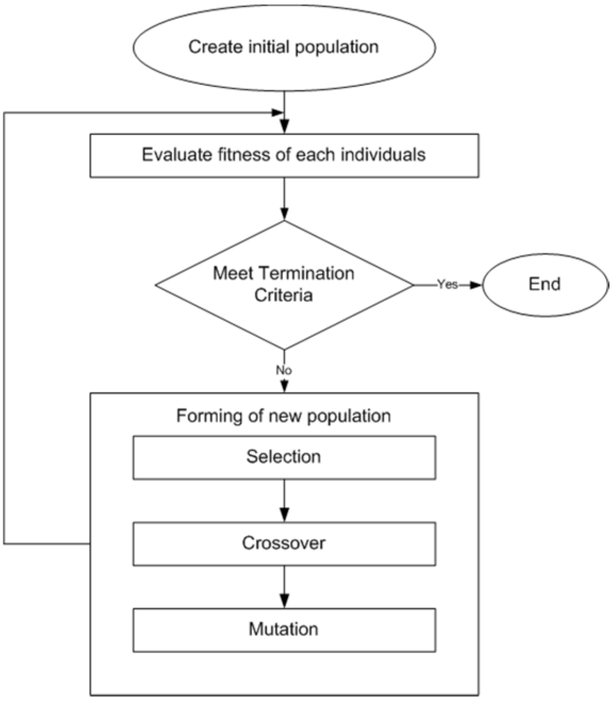
Create initial population
Evaluate fitness of each individuals
Meet Termination Criteria
Yes
End
No
Forming of new population
Selection
Crossover
Mutation

Figure 2 Genetic Algorithm Flowchart. Forming a new generation population of solutions from those selected through a combination of Crossover and Mutation operators.

## 3. Results and Discussion

Liver fibrosis was staged (F0–F4) and required laboratory tests were performed for a cohort of 39567 HCV patients. A group of 33549 patients have mild to moderate fibrosis, and the other group of 6018 patients have advanced fibrosis stage. Patients were divided according to random uniform sampling into two separate sets. About two thirds of the dataset were used for training (n = 22,690 patients) and the rest of data for test (n = 16,877 patients). Mild to moderate fibrosis strata distributed as 19349 patients in the training set and 14200 patients in the test set. Advanced fibrosis strata distributed as 3341 patients in the training and 2677 patients in the test set.

Feature selection was used in our models' constructions to eliminate the redundant features for models simplicity in order to make them easier to interpret by users.

There are three categories of feature selection: wrappers, filters and embedded methods [29]. Filter method for feature selection was used as pre-process before the learning algorithm. It is particularly effective in computation time and robust to over-fitting [30]. Filter method based on Pearson correlation coefficient and scores of significance test for variable ranking was used in this paper.

Table 1 states the characteristics of patients in training and test datasets, and states the P-value and Pearson correlation coefficients between each variable and fibrosis in training set. Data expressed as mean ± SD unless otherwise was stated.

Age, aspartate aminotransferase (AST), platelets count, and albumin were found to be as independent predictors of fibrosis, with highest statistically significant relationship (P-value < 0.0001) and accepted correlation (|r|>0.1) with fibrosis. Therefore, these variables have been used as markers for predicting of advanced fibrosis in proposed models.

In the first proposed model, alternating decision tree was learned for the training data set considering the four variables (which are statistically significant relationship (P-value < 0.0001) and accepted correlation coefficients (|r|>0.1) with fibrosis): age, platelet count, albumin, and aspartate aminotransferase (AST). A tree of 125 nodes was produced in 19 second. Figure 1 (B) shows the decision tree diagram of model 1. In figure 1(B), advanced fibrosis was considered as the positive, while moderate or mild fibrosis was considered as negative. The liver fibrosis of the patient is score by summing all of the prediction nodes through which it passes. You can predict the patient fibrosis stage by follow the represented steps in the section 2.5 in this paper on the proposed ADT model. If the final score of the tree is >= zero (positive value), the patient is high risked to have advanced fibrosis, and vise a versa.

According to the four effective variables, the fibrosis state could be predicted according to the linear equation (5).

$$fibrosis = a_0 + a_1 * Age + a_2 * AST + a_3 * Albumin + a_4 * Platelet \quad (5)$$

Where the parameters' coefficients ($a_0$, $a_1$, $a_3$, $a_4$, and $a_5$) can be optimized by many optimization techniques. In this paper, we compared the performance of PSO, GA, and multi-linear regression (MReg) algorithms in optimizing these parameters' coefficients in purpose of achieving acceptable prediction of the risk of advanced fibrosis model.

Table 1 Characteristics of variables in coherent dataset.

| Characteristics | Training Dataset 22690 | Validation Dataset 16877 | Pearson Correlation Coefficients | P-value |
|---|---|---|---|---|
| **Age (yrs)** | 40 ± 11 | 40±10 | 0.26 | < 0.0001 |
| **Gender** | | | -0.03 | 0.008 |
| **Female** | 6186 (27.3%) | 4555 (26.9%) | | |
| **Male** | 16504 (72.7%) | 12322 (73.1%) | | |
| **BMI** | 26.70 ± 3.79 | 26.79 ± 3.84 | 0.10 | < 0.0001 |
| **AFP (U/L)** | 7.26 ± 26.61 | 7.69 ± 28.49 | 0.10 | < 0.0001 |
| **ALP (U/L)** | 105.41 ± 65.17 | 105.41 ± 65.17 | 0.02 | 0.008 |
| **AST (U/L)** | 57.27 ± 33.73 | 56.78 ± 34.61 | 0.12 | < 0.0001 |
| **ALT (U/L)** | 61.84 ± 36.89 | 61.84 ± 38.19 | 0.06 | 0.008 |
| **Platelet count (*10^9 /L )** | 212.48 ± 60.64 | 211.55 ± 60.86 | -0.18 | < 0.0001 |
| **Albumin (g/dL)** | 4.39 ± 0.42 | 4.40 ± 0.42 | -0.14 | < 0.0001 |
| **Indirect Bilirubin ( mg/dL)** | 0.57 ± 1.77 | 0.60 ± 2.24 | -0.00 | 0.088 |
| **Total Bilirubin (mg/dL)** | 0.76 ± 0.28 | 0.76 ± 0.28 | 0.05 | < 0.0001 |
| **Glucose (mg/dL)** | 96.57 ± 19.41 | 96.69 ± 20.71 | 0.08 | < 0.0001 |
| **Hemoglobin( Hb )** | 14.03 ± 1.47 | 14.03 ± 1.62 | -0.00 | 0.0005 |
| **WBC (10^9/L )** | 6.44 ± 1.90 | 6.44 ± 1.94 | -0.02 | 0.0001 |

Table 2 Parameters' coefficients according different techniques.

| | $a_0$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ |
|---|---|---|---|---|---|
| **GA** | 0.828 | 1.716 | 0.02 | -1.757 | -0.155 |
| **PSO** | 0.013 | 0.1429 | 0.0153 | -1 | -0.0141 |
| **MReg** | 0.224 | 0.007 | 0.001 | -0.058 | -0.001 |

Using PSO, GA, and MReg in optimizing the parameters' coefficients $a_0$, $a_1$, $a_3$, $a_4$, and $a_5$; by training them to give the best ROC values; was outcome a different coefficients' values that stated in table 2. The difference in coefficients' values led to the difference in the accuracy of each algorithm to other.

Modified-PSO algorithm was run on the training dataset to build the second model, with 100 randomly generated individuals (number of particle in the search space) for 100 iterations. Minimum inertia weight is 0.4, and maximum one is 0.9. Cognitive constant $C_1$ (individual learning rate) and social constant $C_2$ are equal to 2.

GA was run on the same cohort in the third model. The present study involves a simple GA formulation involving crossover, mutation and replacement operators. For GA parameters, the following values are used: population size N = 20, maximum number of generations T = 100, crossover fraction = 0.8 using scattered crossover, binary tournament selection operator is used. If the objective value is not improved over 50 generations the algorithm stops. The fitness function is given by calculate the area under the ROC curve using instance parameters in order to maximize it.

The fourth model was built using Multi-linear regression to model the relationship between an advanced fibrosis and the four independent variables: age, AST, albumin, and platelets.

Table 3 states the accuracy, Receiver operating characteristic curve ROC analysis, sensitivity and specificity, positive and negative predictive values of the proposed models for predicting advanced fibrosis on the test set. The machine learning algorithms under study predicted advanced fibrosis in patients with HCC with area under the receiver operating characteristic curve (AUROC) ranging between 0.73 and 0.76 and accuracy between 66.3% and 84.4%. PSO model achieved the highest correlation coefficient with presence of advanced fibrosis, while ADT model achieved the highest accuracy of 84.4% and highest AUROC of 0.75. However, ADtree model shows the lowest NPV of 85.2%, in the time that the rest of proposed models have NPV of 92.2%.

The low sensitivity 7% of the ADT model can be attributed to the zero cut-off frequency, which had been selected by ADT algorithm. The ADT algorithm trained at cut- off point zero. We can choose any other cut-off points from the ROC curve to increase the sensitivity, but it will be at the expense of the accuracy. At Youden index of -0.92 for ADTree model, the model achieved sensitivity of 73% and specificity of 65% but the accuracy down to 66.3% while it was 84.4% using zero cut-off frequency.

The ROC curves for proposed models were plotted in figure 3. The area under ROC for PSO, GA, MReg and ADT were 0.73, 0.75, 0.75, and 0.76, respectively. As demonstrated by ROC curves, it is figured that all proposed model have a close area under ROC, where ADT model has the higher AUROC = 76%, while MReg and GA models have AUROC = 75%, and PSO model has the lowest AUROC =73%.

A cross-validation with 10-fold was used to avoid over training problem. When we applied alternating decision tree algorithm (ADT) on cohort data with the four effective variables using 10-fold cross validation, it achieved 0.75 ROC and 85% accuracy. Multi-linear regression achieved 0.75 ROC and 67.3% accuracy, which is very close to the results of using training and test sets separately.

To examine the effectiveness of Pearson correlation as a filter method to eliminate the redundant features, we used backward elimination technique as a pre-processing step before multi-linear regression algorithm. We run the multi-linear regression algorithm using the fourteen features reported in table 1, and then we started to eliminate the features one by one according to their Pearson correlation coefficients in ascending order. The feature that has the lowest correlation was eliminated first, and so on. It is apparent from the results that using the total number of features did not decrease the accuracy and only very slightly increased the ROC by 2%. This justifies our approach in filtering the features before classification. For further details of the results of backward elimination method referred to in this section, see Appendix 1.

Table 3 Accuracy, sensitivity, specificity, negative and positive predictive values, and ROC analysis of proposed models for predicting advanced fibrosis on test set. Criteria points stated as Youden index for all proposed models except ADT. ADT analysed for two criteria points, first as Youden index, and second for zero value where the ADT trained on.

| Model | Sensitivity | Specificity | Criteria | PPV | NPV | Correlation | ROC | Accuracy |
|---|---|---|---|---|---|---|---|---|
| | % | % | point | % | % | Coefficients | | % |
| **PSO** | 70.4 | 65.6 | >2.715 | 27.9 | 92.2 | 0.29 | 0.73 | 66.4 |
| **GA** | 68.9 | 69.7 | >49.189 | 30 | 92.2 | 0.27 | 0.75 | 69.6 |
| **MReg** | 69.0 | 69.1 | > 0.135 | 29.7 | 92.2 | 0.28 | 0.75 | 69.1 |
| **ADT** | 73.0 | 65.0 | >-0.92 | 28.2 | 92.7 | 0.28 | 0.76 | 66.3 |
| **ADT*** | 07.0 | 99.0 | > =0 | 57.8 | 85.2 | 0.17 | 0.76 | 84.4 |

Abbreviations: PPV, positive predictive value; NPP, negative predictive value; ROC, receiver-operating characteristic curve; PSO, particle swarm optimization; GA, genetic algorithm; MReg, multi-linear regression; ADT, alternating decision tree.

ADT* model with criteria point of zero as trained on, not at Youden index as the rest.

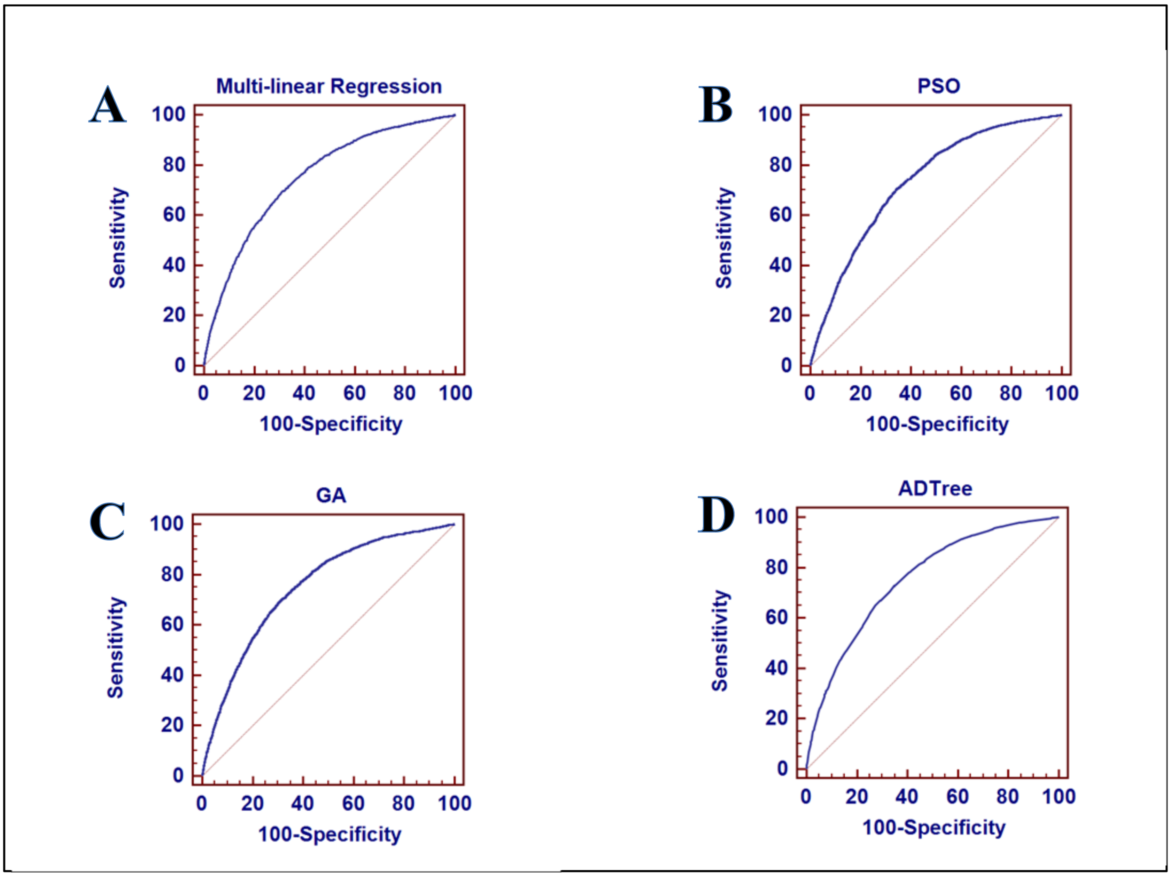

Figure 3 ROC curves plots of proposed models in predicting of advanced fibrosis. (A) ROC curve plot of multi-linear regression model with area under the curve = 0.76. (B) ROC curve plot of particle swarm model with area under the curve = 0.73. (C) ROC curve plot of genetic algorithm model with area under the curve = 0.75. (D) ROC curve plot of alternating decision tree model with area under the curve = 0.75.

## 4. Conclusion

In this study, we made a comparison between different machine learning approaches on prediction of advanced liver fibrosis in Chronic Hepatitis C patients. Particle swarm optimization, decision tree, multi-linear regression and genetic algorithm models were developed. We concluded that we could predict advanced fibrosis stage for chronic HCV patients using different machine learning approaches with high accuracy. The four parameters (age, AST, albumin and platelets count) found to be the most important features in prediction of the advanced fibrosis as they are statistically have significant relationship (P-value < 0.0001) and accepted correlation coefficients (|r|>0.1) with presence of advanced fibrosis as shown in the results. PSO model achieved the highest correlation coefficient 0.28 with presence of advanced fibrosis, while ADT model achieved the highest accuracy of 84.4% and highest AUROC of 0.76. The proposed models could be used as an acceptable, safe, and low cost alternating for predict advanced fibrosis rather than relatively risky alternative tools (such as the liver biopsy) in chronic hepatitis C virus patients.

## Acknowledgement

We thank the Egyptian National Committee for Control of Viral Hepatitis for the patients' information.

## Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## References

[1] Laurent Castera. Noninvasive methods to assess liver disease in patients with hepatitis B or C. GASTROENTEROLOGY. 2012; 142:1293–1302.

[2] 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. 2012; 12:55.

[3] Mahmoud ElHefnawi, Mahmoud Abdalla, Safaa Ahmed, Wafaa Elakel, Gamal Esmat, Maissa Elraziky, Shaima Khamis, et al. Accurate prediction of response to Interferon-based therapy in Egyptian patients with Chronic Hepatitis C using machine-learning approaches. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2012; 771-778.

[4] Gravitz L. A smouldering public-health crisis. Nature474 (7350). 2011.

[5] Abdulaziz Q. Alodini. Prevalence of Hepatitis B Virus (HBV) and Hepatitis C Virus (HCV) Infections among Blood Donors at Al-Thawra Hospital Sana'a City-Yemen. Yemeni Journal for Medical Sciences. 2012.

[6] Ehab Nashaat. Lipid profile among chronic hepatitis C Egyptian patients and its levels pre and post treatment. Nature and Science. 2010; 8(7).

[7] Dana Crisan, Corina Radu, Mircea Dan Grigorescu, Monica Lupsor, Diana Feier, Mircea Grigorescu. Prospective non-invasive follow-up of liver fibrosis in patients with chronic hepatitis C. J Gastrointest Liver Dis. 2012; pp. 375–382.

[8] Pierre Bedossa, Fabrice Carrat. Liver biopsy: the best, not the gold standard. J Hepatology. 2009; 50: 1-3.

[9] Arie Regev, Mariana Berho , Lennox J Jeffers, Clara Milikowski, Enrique G Molina, Nikolaos T Pyrsopoulos, Zheng-Zhou Feng, et al. Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection. Am J Gastroenterology. 2002; 97: 2614- 2618.

[10] Hugo Rosen. Clinical Practice. Chronic Hepatitis C Infection. The New England Journal of Medicine. 2011; 364 (25): 2429–38.

[11] Parkes J, Guha IN, Roderick P, Rosenberg W. Performance of serum marker panels for liver fibrosis in chronic hepatitis C. J Hepatol. 2006; 44:462-474.

[12] Moon Young Kim, Woo K young Jeong, Soon Koo Baik, Invasive and non-invasive diagnosis of cirrhosis and portal hypertension, World J Gastroenterology. 2014 April 21; 20(15): 4300-4315.

[13] Somaya Hashem, Gamal Esmat, Wafaa Elakel, Shahira Habashy, Safaa Abdel Raouf, Samar Darweesh, Mohamad Soliman, Mohamed Elhefnawi, Mohamed El-Adawy, and Mahmoud ElHefnawi. "Accurate Prediction of Advanced Liver Fibrosis Using the Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients," Gastroenterology Research and Practice, Volume 2016 (2016).

[14] Bonacini M, Hadi G, Govindarajan S, Lindsay KL. Utility of a discriminant score for diagnosing advanced fibrosis or cirrhosis in patients with chronic hepatitis C virus infection. Am J Gastroenterology. 1997; 92: 1302-1304.

[15] Somaya Hashem, Shahira Habashy, Wafaa El-Akel, Safaa Abdel Raouf, Gamal Esmat, Mohamed El-Adawy & Mahmoud El-Hefnawi. A Simple multi-linear regression model for predicting fibrosis scores in chronic Egyptian hepatitis C virus patients. International Journal of Bio-Technology and Research (IJBTR). 2014 Jun; Vol. 4, Issue 3, 37-46.

[16] Richard K. Sterling, Eduardo Lissen, Nathan Clumeck, Ricard Sola, Mendes Cassia Correa, Julio Montaner, Mark S. Sulkowski, et al. Development of a simple non-invasive index to predict significant fibrosis in patients with HIV/ HCV co-infection. Hepatology. 2006; 43:1317–25.

[17] Chun-Tao Wai, Joel Greenson, Robert Fontana, John Kalbfleisch, Jorge Marrero, Hari Conjeevaram, and Anna Lok. A simple non-invasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology. 2003; 38:518–26.

[18] Foucher J, Chanteloup E,Vergniol J, Castéra L, Le Bail B, Adhoute X, Bertet J, et al. Diagnosis of cirrhosis by transient elastography (FibroScan): a prospective study. Gut. 2006; 55: 403–408.

[19] Gaia S, Carenzi S, Barilli AL, Bugianesi E, Smedile A, Brunello F, Marzano A, et al. Reliability of transient elastography for the detection of fibrosis in non-alcoholic fatty liver disease and chronic viral hepatitis. J Hepatol. 2001; 54: pp. 64–71.

[20] Ashraf Wahba, Nagat Mohammed, Ahmed Seddik, Mohamed El-Adawy. Liver fibrosis recognition using multi-compression elastography technique. Journal of

Biomedical Science and Engineering JBiSE. 2013; Vol.6 No.11, 1034-1039.

[21] Bedossa P., Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology.1996; 24:289–293.

[22] Danan Wang, Qinghui Wang, Fengping Shan, Beixing Liu, Changlong Lu. Identification of the risk for liver fibrosis on CHB patients using an artificial neural network based on routine and serum markers. BMC Infectious Diseases. 2010; 10:251.

[23] Leo Breiman, Jerome H Friedman, Richard A Olshen, Charles J Stone. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. 1984.

[24] Quinlan J. The Proceedings of the 30th National Conference on Artificial Intelligence. Menlo Park, CA. Bagging, boosting, and C4.5. 1996; pp. 725–730.

[25] Freund Y., Mason L. The alternating decision tree learning algorithm. Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia. 1999; 124-133.

[26] http://en.wikipedia.org/wiki/Alternating_decision_tree.

[27] Sivanandam S.N., Deepa S. N. Introduction to Genetic Algorithms. ISBN 978-3-540-73189-4 Springer Berlin Heidelberg New York 2008.

[28] Kennedy, J., and Eberhart, R. C. (1995). Particle swarm optimization. Proc. of IEEE International Conference on Neural Networks (ICNN), Perth, Australia, 1995; Vol.IV, pp.1942-1948.

[29]Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). An Introduction to Statistical Learning. Springer. p. 204.

[30] https://en.wikipedia.org/wiki/Feature_selection.

Somaya Hashem, MSc, BSc.
Assistant Researcher at Department of Information and Systems.
Engineering Devision.
National Research center, Cairo, Egypt.

Gamal Esmat, MD
Professor of Hepatology and Endemic Medicine – Cairo University.
Vice-President of Cairo University for graduate studies and research.
State award from Academy of Scientific research, Egypt in the field of Medical science 2010.
Past President of International Association for study of liver diseases 2006 – 2008.
Member of WHO strategic and advisory committee for viral hepatitis.
Founder of Viral Hepatitis Treatment Center MOH.
Founder of Middle East School of Hepatology.

Wafaa Elakel, MD
Associate Professor at Department of Endemic Medicine and Hepatology, Faculty of Medicine, Cairo University, Cairo, Egypt.

Shahira Habashy, PhD, MSc, BSc
Associate Professor at Communications, Electronics and Computers Department, Faculty of Engineering, Helwan University, Cairo, Egypt.

Safaa Abdel Raouf, PhD, MSc, BSc.
Associate Professor at Department of Information and Systems. Engineering Devision. National research center, Cairo, Egypt.

Mohamed Elhefnawi, PhD, MSc, BSc.
Communications and Computer Department, Faculty of Engineering, Modern University, Cairo, Egypt.

Mohamed El-Adawy, PhD, MSc, BSc.
Professor at Communications and Computer Department, Faculty of Engineering, Modern University, Cairo, Egypt.

Mahmoud ElHefnawi, PhD, MSc, BSc
The biomedical informatics and Chemoinformatics group leader at the Center of Excellence for Advanced Sciences (CEAS) and Associate Professor at the Informatics and Systems Department, National Research Center (NRC). He is also a part-time faculty/ Senior Research scientist at Nile University and was so at the Egyptian-Japanese University for Science and Technology (EJUST), and had affiliations/ part-time participations/ consultations at the Youssif-Jamil Science and Technology Research Center (YJ-STRC) and was a part-time adjunct faculty at the American University in Cairo teaching graduate level courses in Bioinformatics.