Prediction of Optimum Treatment in Beta-Thalassemia Patients with Hepatitis C Virus Infection

Aisha M. Hussein[1], Ahmed Sharaf Eldin[2], Sanaa M. Kamal[3], A. Abdo[4], Samar K. Kassim[5], Khaled A. Bahnasy[6]

[1]*Faculty of Computers and Information, Helwan University, Cairo, Egypt*
[2]*Faculty of Computers and Information, Helwan University, Cairo, Egypt*
[3]*Faculty of Medicine, Ain Shams University, Cairo, Egypt*
[4]*Faculty of Computers and Information, Helwan University, Cairo, Egypt*
[5]*Faculty of Medicine, Ain Shams University, Cairo, Egypt*
[6]*Faculty of Computers and Information, Ain Shams University, Cairo, Egypt*
[1]*Corresponding Author: Aisha M. Hussein*

Abstract

Aim: Iron overload and hepatitis C virus (HCV) infection together can lead to chronic liver damage in β-thalassemia major (β-TM) patients. Given the high toxicity of this regimen, Peginterferon plus ribavirin (peg-IFN/RBV) is still the standard of care for treatment of HCV in many countries. These considerations led us to design a prediction model in combination of data mining techniques to compare treatment efficacy and tolerability of Peg-IFN and RBV combination therapy versus Peg-IFN monotherapy.

Materials and Methods: Data from (β-TM) patients with genotype-GT4 from different centers in Egypt were analyzed. Treatment endpoint is sustained virological response (SVR) defines as undetectable HCV RNA 24 weeks following termination of therapy. The prediction models were produced with decision tree classification algorithms: J48 Tree and Logistic Model Tree (LMT).

Results: A total of 175 patients who completed Peg-IFN/RBV therapy achieved SVR at 24 weeks 87(49.7% responders) and 88 did not (50.3%). While 135 patients completed peg-IFN monotherapy treatment achieved 59(43.7% responders) and 76(56.3% non- responders). The accuracies of compared models are above 70%. In Peg-IFN/RBV, LMT algorithm shows 88.2% accuracy compared to 82.4% by J48 model. In Peg-IFN monotherapy, LMT shows 69.2% in comparison to 82.4% by J48, J48 is much simpler and easier to interpret. We investigated the role of iron overload on the efficacy of anti-HCV treatments. Various cutoff levels of Ferritin were significantly related to different probability of SVR. Finally, we evaluated the changes in blood transfusion regime, laboratory and histopathological data during the antiviral treatment. This was further confirmed by binary regression analysis; p value<0.05.

Conclusion: The experimental results showed that the model obtained acceptable results.

Key Word: *Data Mining, β-TM, HCV, SVR, Decision Tree.*

I. INTRODUCTION

Thalassemias are the commonest monogenic disorders in the world [1], and the incidence rate is higher in the Middle East [2]. β-thalassemia constitutes a major health problem in Egypt with an estimated carrier rate of 9-10% [3]. Beta thalassemia major, also called Cooley's anemia is prevalent in the Mediterranean regions particularly Italy, Greece, Egypt. It is also found in Saudi Arabia, Iran, Africa, Southeast Asia and southern China [4].

The thalassemia disease classified into three categories: Thalassemia Major (Thal-M), Thalassemia Intermediate (Thal-I), and Thalassemia Trait or Minor (Thal-T). Patients with Thal-M and Thal-I require repeated blood transfusions during whole their lives and other treatments, while Thal-T usually do not require any specific therapy [5]. Since thalassemia is a genetic disease so we can control it by eliminate the marriage between the both people who have carried the genes. Complete Blood Count (CBC) test is the first and the simplest test can perform on the way to indicate the thalassemia's genes existence [6].

Transfusion-related infections especially chronic HCV infection increases the morbidity and mortality of thalassemic patients. Despite remarkable improvements in HCV treatment during the recent years, many thalassemic patients with chronic HCV infection still do not respond adequately to standard therapy [7].

In thalassemic patients, monotherapy (Peg-IFN) with or without ribavirin has been the available treatment for decades. [7,8,9] Ribavirin, that is one of the major determinants of the SVR, can induce life-threatening anemia in thalassemia major patients, and thus, is generally considered contraindicated in these patients [10,11]. In addition to contraindication of ribavirin, high iron content of serum and liver with its probable role in conferring patients' interferon resistance and its synergistic effect on progression of liver fibrosis is another major confounder in treatment of thalassemia major patients [12].

Extensive analysis has identified several factors that correlate with (SVR) such as gender, age, viral load and degree of fibrosis on liver biopsy. Most complications can be avoided if ferritin levels can be brought down to <1500 μg/l. Several of these predictive factors may be helpful clinically in guiding therapy, in providing advice about the likelihood of a response and in determining the duration of therapy [13,14].

Few HCV GT4-infected patients were enrolled in registration trials for PegIFN alfa/RBV [13]. A comprehensive review of trials in GT4 patients, most of which were conducted in the Middle East, found that (SVR) rates have ranged from 38% to 87% [13]. The SVR rate of 87% was obtained in HCV GT4-infected patients with a rapid virologic response (RVR), defined as undetectable HCV RNA at week 4 of treatment [14]. Consistent with this finding, an SVR rate of 86% was obtained in Egyptian patients with HCV GT4 infection who achieved an RVR [16]. In a study performed in France, an RVR was obtained in 15 out of 59 patients with HCV GT4 infection and was a strong predictor of SVR with 82%, 47%, and 29% of patients with host *IL28B* rs12979860 CC, CT, and TT genotypes, respectively, achieved an SVR [17].

The ability to predict which patients are most likely to achieve an SVR with PegIFN alfa-2a/RBV on the basis of patient characteristics alone would be extremely useful in guiding treatment decisions. Moreover, the combination of a baseline prediction tool and on-treatment predictors such as RVR would allow clinicians to optimize treatment outcomes by preferentially initiating treatment in patients with favorable baseline profiles and continuing treatment only in those most likely to respond, while avoiding exposure to potential adverse events in those unlikely to achieve an SVR.

Decision-tree models represent a potential tool for prediction of therapeutic outcome to antiviral therapy among difficult-to-treat GT4 chronic HCV patients, as they enable clinical decisions either to estimate the probability of patients will develop a disease, will respond to a treatment, their disease will recur or regarding the proper selection of patients for therapy without imposing any additional costs [18].

A combination of a tree structure and logistic regression models (i.e. a decision tree with logistic regression functions at the leaves) resulting in a single tree. An advantage of using logistic regression is that explicit class probability estimates are produced rather than just a classification. In this paper, we present a method, called LMT (Logistic Model Trees) that follows this idea.

II. RELATED WORK

In a study, Wongseree et.al [16] investigated thalassemia classification by using a neural network and a decision tree, which is evolved by genetic programming. The aim is to differentiate between thalassemic patients, persons with thalassemia trait and normal subjects by inspecting characteristics of red blood cells, reticulocytes and platelets. But they need in the proposed model more blood testing like Platelet and Reticulocyte. El-Sebakhy and Elshafei in [17] proposed thalassemia screening using unconstrained functional networks classifier and compare the performance of the proposed model with both multilayer perceptron (MLP) and support vector machine (SVM), and the results showed that using unconstrained functional networks classifier takes much less computations. Zayed al., [18] compiled data from Egyptian patients with chronic HCV-G4 and used Weka implementation C4.5 decision-tree algorithm and Reduced Error Pruning (REP) tree as one of the Fast decision tree learner (FDTL) algorithms and they concluded that the Prediction of treatment outcome in chronic HCV patients genotype-4 (HCV-G4) has been an important debate since even with the application of combination therapy for 48 weeks only around 50% of patients will respond.

On studying challenges of diagnosis of iron deficiency anemia (IDA), β-thalassemia trait and α-thalassemia trait to help specialists to detect the type of anemia accurately, regarding their blood similarities and number of the patients, avoid doing irrelevant tests, reduce time and the price of detection only using the CBC test. This study showed by modeling J48, IBK and Naïve Bayes using hybrid (voting) algorithm had the highest performance and low mean absolute error than using single classifier including Multi-layer perceptron [19].

Previously in a meta-analysis of literature, showed that peg-IFN monotherapy can induce a SVR rate of 28% in thalassemia major patients, while peg-IFN/Ribavirn (low dose therapy, accompanying by tight monitoring of patients) elicits a SVR rate of 44 percent [20]. The small sample size and lack of control group in included studies were two major downsides of meta-analysis.

Indeed, this study has been designed to address this issue in a model and to reach a more precise estimation of ribavirin tolerability and the effect of its low dose administration on improving SVR in HCV infected thalassemia major patients.

III. PATIENTS AND METHOD

- **Clinical Data**

Data from 310 β-TM patients infected with HCV-G4 from different centers in Egypt were analyzed. From them, group 1 (n=175) who received Peg-IFN/Ribavirin therapy and group 2 (n= 135) with Peg-IFN monotherapy. All patients were evaluated at baseline and at different time points during treatment and follow up. The treatment endpoint was SVR defines as undetectable HCV RNA 24 weeks during or following termination of therapy. The dataset created contains demographic variables for the patients have been investigated and analyzed.

Table 1 states relationship between baseline independent variables and SVR at 24-week with statistically significant relationship at (P value < 0.05). Data were reported also as mean value ±standard deviation (SD).

Table (1): Baseline Characteristics of Dataset Attributes

Features name	Group1	p-value	Group2	p-value
Patients No.	175		135	
Age (Years)		0.19		0.27
(Means ± SD)	26.0 ±4.96		25.7± 4.7	
Range	18-34		18-34	
Gender		0.35		0.23
Male (1)	122 (69.7%)		98 (72.6%)	
Female (2)	53 (3.3%)		37 (27.4%)	
BMI , Mean ± SD	26.73 ±2.25	0.18	26.9± 2.68	0.86
Range (16.6-43.2)	23-32		22-31	
White blood cells		0.72		0.59
(4000-11,000) Range	2600-11150		3904-11173	
Mean ± SD	6433.6 ±20516		7626.2 ±2148.4	
Hemoglobin (mg/dl)		0.61		0.42
Range	6-12		6-12	
M(13-16), Mean ± SD	(122) 9.1 ±0.189		(98) 9.03 ±2.01	
F (12-14), Mean ± SD	(53) 9.06±0.27		(37) 9.1±1.94	
Platelets		0.54		0.85
Range (150,000-450,000)	174,426-224,985		174,334-224,854	
Mean ± SD	2004,457 ±138,1		200680.9 ±147,58	
Baseline PCR (IU/ml)				
Mean ± SD	465844.7 ±550732.1	0.10	1051113.1±277355.2	0.42
Range	98700-7049083		148317-2380000	
<600,000	123 (70.3%)		5 (3.7%)	
>600,000	52 (29.7%)	0.01*	130 (96.3%)	0.014*
AST (U/L), Mean ± SD	90.41 ±22.3	0.11	91.39 ± 23.4	0.15
Range	49-127		49-128	
Normal (<35 U/L), No.%	60 (34.3%)		46 (34.1%)	
Elevated No. (%)	115 (65.7%)		89 (65.9%)	
(ALT), Mean ± SD	86.6 ±22.4	0.18	86.4 ±21.9	0.86
Range	49-128		50-127	
Normal (<35 U/L), No.%	74 (42.3%)		56 (41.5%)	
Elevated No. (%)	101 (57.7%)		79 (58.5%)	
Histologic_staging (1-6)	1-3	0.78	1-5	0.21
Mean ± SD	1.99 ±0.82		2.37± 1.15	
Histologic_grading		0.57		0.48
(Ishak score 1-18)	4-13		4-13	
Mean ± SD	8.71 ±2.61		8.61± 2.8	
Stage of liver siderosis(1)	40 (22.8%)	0.33	32 (23.7%)	0.17
(2)	36 (20.5%)		36 (26.6%)	
(3)	48 (27.4%)		32 (23.7%)	
(4)	51 (29.1%)		35 (25.9%)	
Transfusion requirement %				
Yes (1)	98 (56.0%)	0.25	6 (4.4%)	0.53
No (2)	77 (44.0%)		129 (95.6%)	
Ferritin before treatment				
Mean ± SD	2066.05 ±543.6		1992.46±552.66	
Range (ng/mL)	1135-2997		1109-3006	
>1560	131 (74.9%)	0.03*	100 (74.8%)	0.62
<1560	44 (25.1%)		35 (25.2%)	
Ferritin after treatment				
Mean ± SD	2101.5 ±527.82		2008.9±536.95	

Range (ng/mL)	1111-3007		1114-3016		
>1560	141 (80.6%)	0.91	98 (72.6%)	0.70	
<1560	34 (19.4%)		37 (27.4%)		
Splenectomy, y,No.(%)		0.12		0.51	
Yes (1)	93 (53.1%)		70 (51.9%)		
No (2)	82 (46.9%)		65 (48.1%)		
IL28B polymorphism		0.78		0.14	
1=CC	47 (26.9%)		40 (29.6%)		
2=CT	62 (35.4%)		51 (37.8%)		
3=TT	66 (37.7%)		44 (32.6%)		
HCV-RNA 4 week (RVR)					
Respond	21 (11.9%)	0.0*	15 (11.1%)	0.0*	
Not respond	154 (87.5%)		120 (88.9%)		
HCV-RNA 24 week (class)					
Respond	87 (49.7%)		59 (43.7%)		
Not respond	88 (50.3%)		76 (56.3%)		

The dataset (age, gender, BMI), hematological variables (white blood cells (WBC), hemoglobin (Hb), platelet count), biochemical variables [aspartate aminotransferase (AST) and alanine aminotransferase (ALT), ferritin levels before and after treatment, quantity of HCV-RNA] in addition to siderosis grade, transfusion requirement, splenectomy, IL28B polymorphism and histopathologic features of chronic hepatitis.

- **Predictor of SVR**

The pre-treatment clinical data including age, gender, liver histological (Ishak score) stage (1-6) and grade (1-18), hepatocellular siderosis (iron-overload) grade (1-4), increase transfusion requirement, baseline HCV-RNA quantity, AST, ALT, ferritin levels before and after treatment, IL28-β (CC, CT, TT) were elucidated as independent factors in relation to SVR at week 24 and evaluated by binary logistic regression. The overall rate of SVR in group (1) patients was 65.1% while in group (2) was 70.9% (Table 2).

Table (2): Binary logistic regression for baseline factors associated with likelihood of SVR of treatment at (24 weeks).

Parameter	Group 1: Peg-IFN/RBV				Group 2: Peg-IFN			
	Odds ratio	SE	95% C.I.	P value	Odds ratio	SE	95% C.I.	P value
Age	1.044	0.03	(0.978-1.115)	0.197	0.96	0.04	(0.88-1.04)	0.35
Male gender	1.026	0.36	(0.501-2.10)	0.94	0.77	0.43	(0.33-1.8)	0.56
Hb (mg/dl)	1.06	0.08	(0.9-1.2)	0.23	0.95	0.09	(0.78-1.16)	0.66
Histologic_grade	1.0	0.06	(0.8-1.1)	0.89	0.99	0.07	(0.86-1.1)	0.9
Histologic_stage	1.1	0.2	(0.7-1.6)	0.9	0.99	0.17	(0.65-1.3)	0.67
Siderosis_grade	0.8	0.15	(0.6-1.1)	0.3	1.0	0.17	(0.7-1.4)	0.99
Transfusion (Yes)	1.4	0.33	(0.7-2.7)	0.2	1.4	0.9	(0.23-9.1)	0.67
Spleenectomy (Yes)	0.7	0.3	(0.3-1.4)	0.4	1.1	0.39	(0.5-2.5)	0.68
Baseline_HCV-RNA (>600 IU/ml)	1.0	0.39	(1.0-1.0)	0.02*	1.0	0.0	(1.0-1.0)	0.0*
AST (>100 U/L)	0.98	0.35	((0.9-1.0)	0.04*	0.85	0.40	(0.38-1.8)	0.69
ALT (>100 U/L)	1.0	0.36	(0.9-1.0)	0.05*	0.66	0.43	(0.28-1.5)	0.34
Ferritin_before (< 1560 ng/ml)	1.0	0.41	(1.0-1.0)	0.03*	0.65	0.47	(0.25-1.66)	0.37
Ferritin_after (< 1560 ng/ml)	1.0	0.42	(0.9-1.0)	0.2	0.74	0.45	(0.3-1.8)	0.51
IL28β	0.8	0.4	(0.3-1.8)	0.6	0.76	0.5	(0.27-2.08)	0.59
Overall prediction (%)	**65.1%**				**70.9%**			

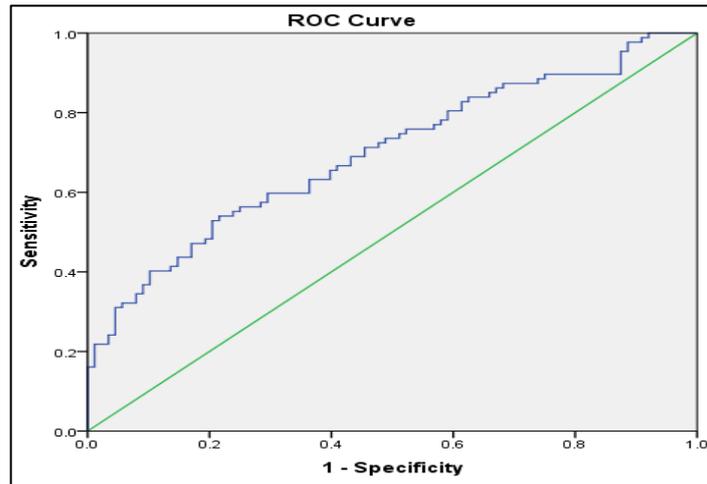Significant at p-value <0.05

Fig (1): shows ROC curve for predicted model of group 1 (Peg-IFN/Ribavirin) with AUC of 0.69, sensitivity 70% and specificity about 45%.

- **Data mining prediction Model**

A framework has been built to predict the response of β-TM with HCV genotype four patients to treatment by applying Data Mining Techniques (DMT) on clinical information, and then extract the result of DMT to be a Knowledge Base for our application to perform the prediction process.

- *Feature Selection and Classification*

To make the untreated data more suitable for further analysis, pre-processing steps should be applied. Some different techniques exist, with these techniques (1) feature selection, (2) dimensionality reduction, and (3) feature extraction. In the classification task, the data set is divided into training and testing sets. A good classification model should be appropriate the training set well and correctly classifies all the examples. If the test error rates of a classification model begin to increase although the training error rates decrease, then this is known as data overfitting.
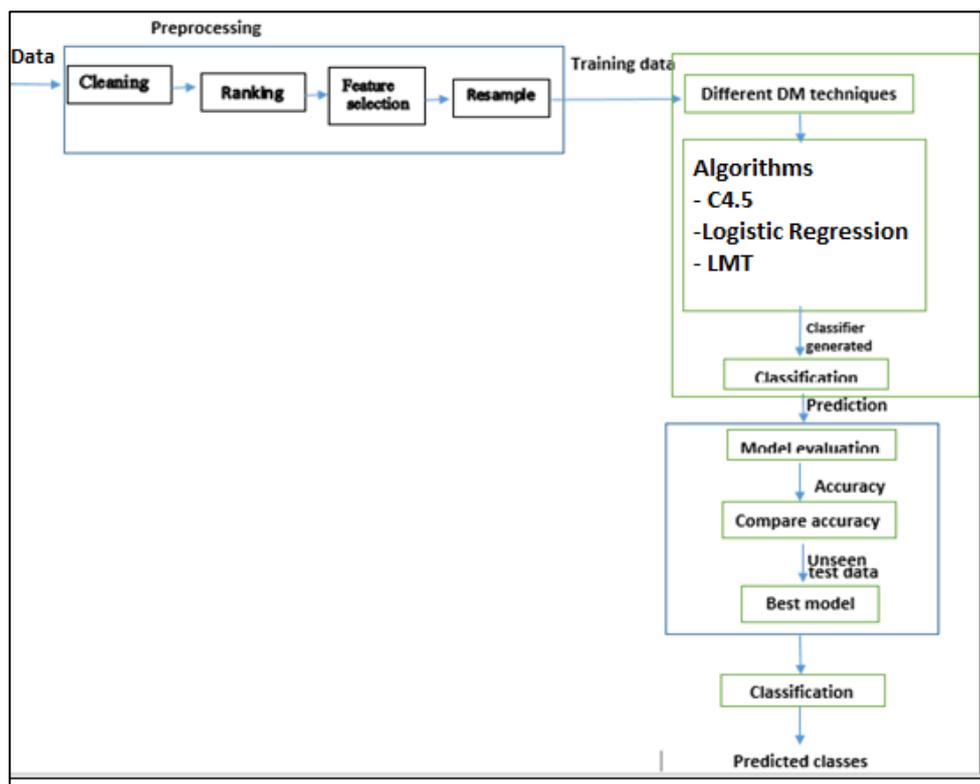
Fig (2): Framework for proposed data mining algorithms

- **The Proposed Model**

First Stage: Data Collection and Preprocessing

At preprocessing dataset: should eliminate useless attributes, refill the missing values and remove/refill the outlier values on the outlier samples. Data sets have been split into the training set, which is used to build the model and testing set that is used to evaluate the proposed model.

Second Stage: Feature selection and reduction

Is to select a subset of features relevant to the target DMT from all the features of the data set. In the filtering approach; the feature selection algorithm is independent of the DMT which applied to the selected features. A subset of 15 features was selected to speed up the model building process. The class label is considered as the PCR of (24th weeks following termination of therapy).

For Resampling; each treatment group has been divided into: 90% of the data as a training set and 10% as a testing set using supervised resample filter in Weka.

Third Stage: Evaluating the Proposed Model

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.

- *Decision Trees*
- <u>**C4.5 (Weka J48)**</u>

C4.5; is based on the ID3 algorithm will emphasize on splitting dataset and selecting a test that will give best result in information gain. If all cases are of the same class, the tree is a leaf and the leaf is returned labeled with this class. For each attribute, calculate the potential information provided by a test on the attribute (based on the probabilities of each case having a particular value for the attribute). Depending on the current selection criterion, find the best attribute to branch on.

Entropy: measure disorder of the data, if the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

Entropy (D) $= -\sum_{i=1}^{c} pi \, log2 \, pi$

The information gain is based on the decrease in entropy after a dataset (D) is split on an attribute, that returns the highest information gain.

Information (after split attribute y into partitions) to classify D:

Entropy Y (D) $= \sum_{j=1}^{v} \frac{|Dj|}{|D|}$ *Entropy (Dj)

Total information gain by branching on attribute Y:

Gain (Y) = Entropy(D) – Entropy Y (D)

Pruning: After the whole creation processes of the tree, which classify all the training set instances, it is pruned. This is to minimize classification errors which can be occurred because of specialization in the training set; we do this to make the tree more general. Replacement is by raising node subtree upwards from leaves toward the root of tree and also replacing other nodes on the same way.

C4.5 Tree model revealed that in patients who were adherent to PegIF treatment, HCV-RNA baseline was selected as the variable of initial split (most decisive) (Fig.4).

Among patients who were adherent to Peg-IFN/RBV treatment, patients with baseline ferritin levels >1679 ng/ml were classified as high probability group (Fig.3).

Of the 310 total patients, end of treatment response at week 24 was evident in group (1) 49.7% treated with (Peg-IFN/RBV) and 43.7% group (2) treated with (Peg-IFN monotherapy). Failure in response and breakthrough were evident in 50.3% group (1) and 56.3% group (2).
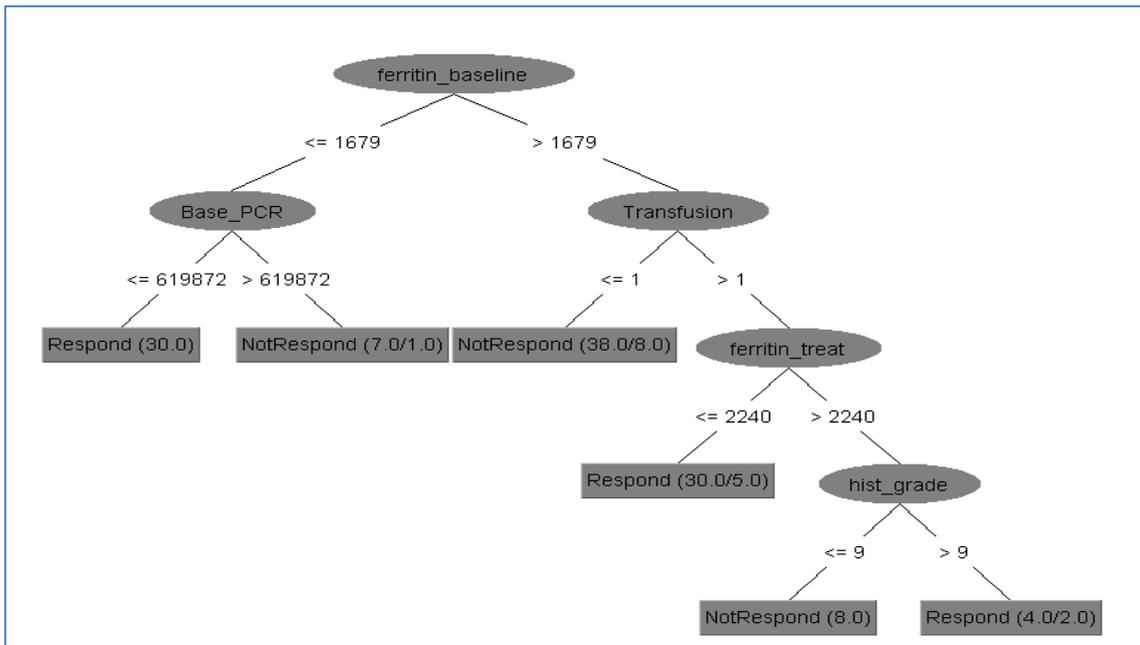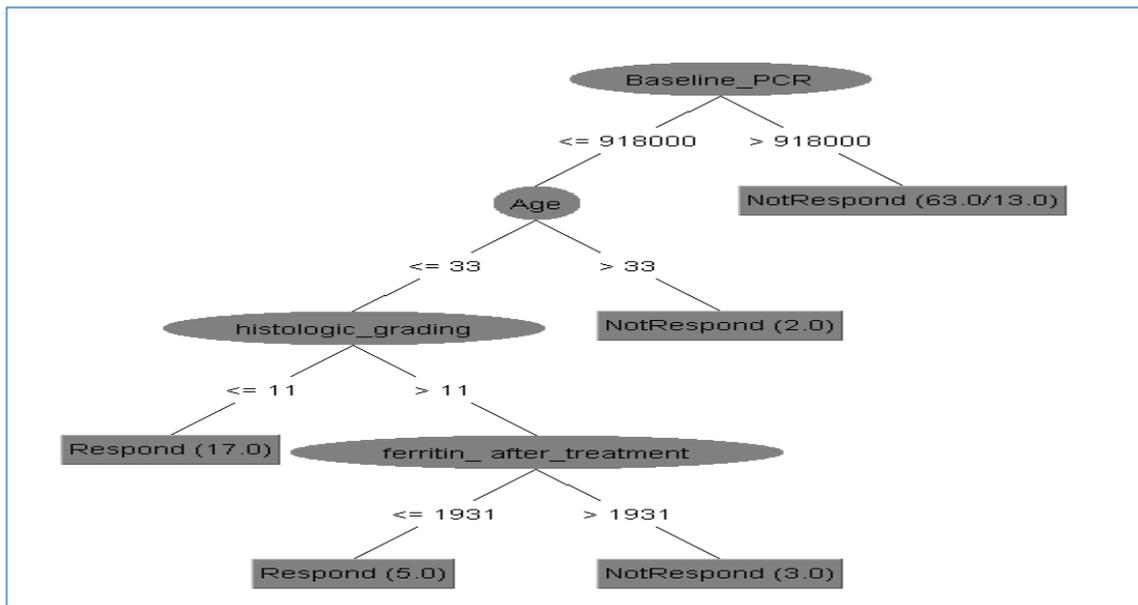
Fig (3): DT for (Peg-IF /RBV) treatment by C4.5 algorithm

Fig(4): DT for Peg-IF monotherapy by C4.5 algorithm

- **Logistic Regression and Logistic Model Tree (LMT)**

The algorithm learns a coefficient for each input value, which are linearly combined into a regression function and transformed using a logistic function. This method seeks to simplify the model during training by minimizing the size of coefficients learned by the model. The logistic regression model can be described as 21]:

P (y=1| X;Ө) that represents the probability of y=1 given X and parameterized with Ө

$$p \ (y_i \mid x_i \ ; \Theta) = \sum_{j=1}^{n} x_j^{(i)} \theta_j = \ \theta^T X \qquad (1)$$

where yi = [0,1], Ө= takes any real number.

Change the probability to ODDS   odds = p/1-p

Log (p / 1-p) = $\theta^T X$ (2)

To estimate parameter Θ;

p (y=1 | x$_i$; Θ) = $\frac{1}{1+e^{-\theta^T xi}}$ (3)

p (y=0 | x$_i$; Θ) = $1 - \frac{1}{1+e^{-\theta^T xi}}$ (4)

Both equation can combine:

p (y $_i$ |x$_i$; Θ) = $(\frac{1}{1+e^{-\theta^T xi}})^{yi} \cdot (1 - \frac{1}{1+e^{-\theta^T xi}})^{1-yi}$ (5)

LMT are classification trees with logistic regression functions at the leaves. LMT significantly produces smaller trees than C4.5 (due to its pruning mechanism).
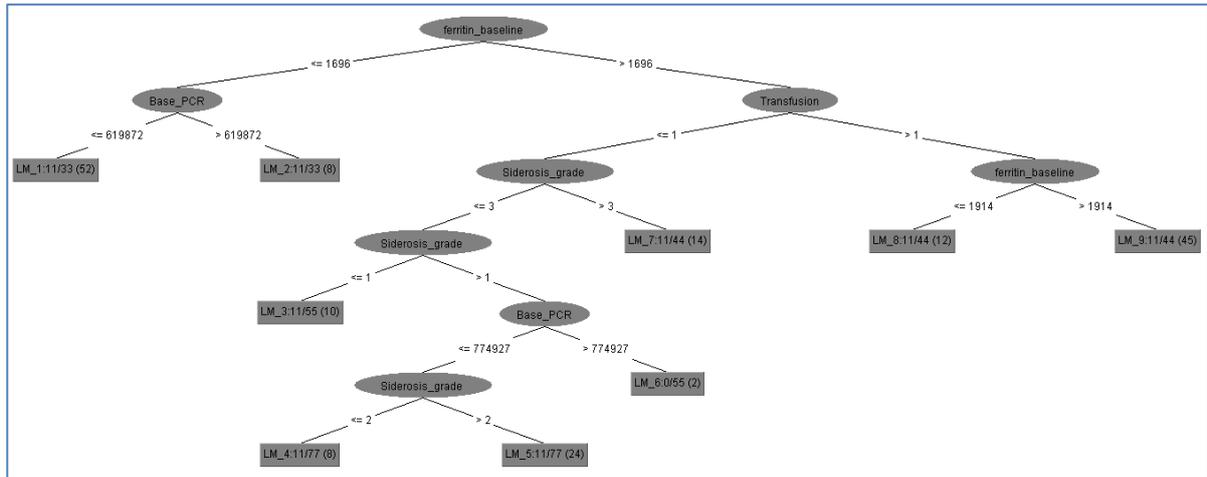
Fig (5): DT showed predicted probability estimate for (Peg-IF/RBV) treatment by LMT algorithm

IV. EXPERIMENTAL RESULTS

Validation of the decision tree and various algorithms:

Was performed with test mode hold-out validation that considered as the simplest type of cross validation method. The data set is divided into two different sets, which are known as the training set and the testing set [22]. The function approximator uses the training set only to fit a function which is used to predict the output values for the testing set data which has never seen these output values before. Then the errors it makes are gathered as before to give the mean absolute test set error, which will be used to evaluate the model. This method is usually preferable to the other methods and takes less time to compute. In this study we classified the accuracy of different algorithms and compared the results to know which algorithm shows best performance.

After applying the model; a large scale of statistical information was obtained. These performance measures had been used to evaluate the model as shown in Tables (3).

Table (3): Performance for group1 (combined therapy) and group2 (monotherapy) after testing

|	Group1 (Peg-IFN/Ribavirin)			Group2 (PegIF)		
	J48	Logistic regression	LMT	J48	Logistic regression	LMT
Total number of instances	175	175	175	135	135	135
Total number of instances after 90% split	17	17	17	13	13	13
Size of the tree	27		17	35		15
Number of leaves	14		9	18		8
Kappa statistics	0.62	0.28	0.74	0.56	0.35	0.41
Mean absolute error	0.19	0.36	0.16	0.19	0.39	0.35
Relative absolute error %	40.3	76.2	34.7	42.9	82.3	78.1
Sensitivity % (TPR)	90.0	77.8	90.9	66.7	40.0	42.9
Specificity % (TNR)	71.4	50.0	83.3	90.0	87.5	100.0
PPV% (Precision)	81.8	63.6	90.9	66.7	66.7	100.0
NPV%	83.3	66.7	83.3	90.0	70.0	60.0
FNR	18.2	36.4	9.1	33.3	33.3	0.0
FPR	16.7	33.3	16.7	10.0	30.0	40.0
ROC area	0.88	0.72	0.86	0.9	0.68	0.77
Accuracy %	82.4	64.7	88.2	84.6	69.2	69.2
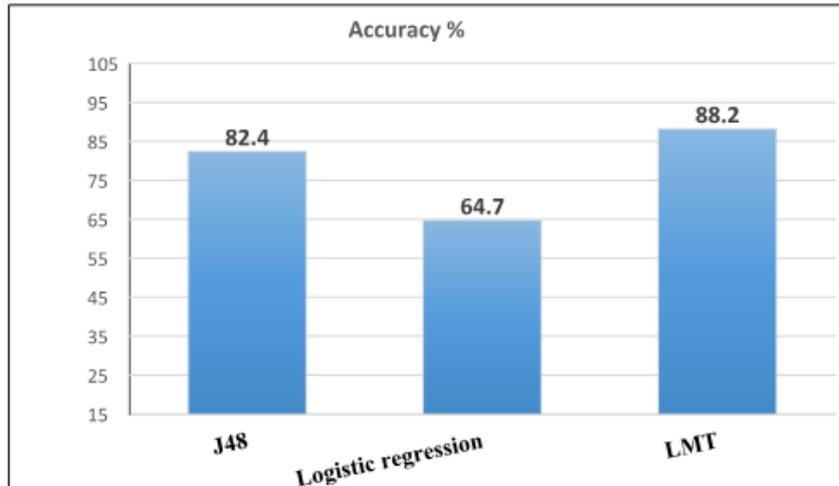
Fig (6): Accuracy% in patients treated with Peg-IFN/Ribavirin using different algorithms. J48 of optimal accuracy at 82.4% and LMT of 88.2%.
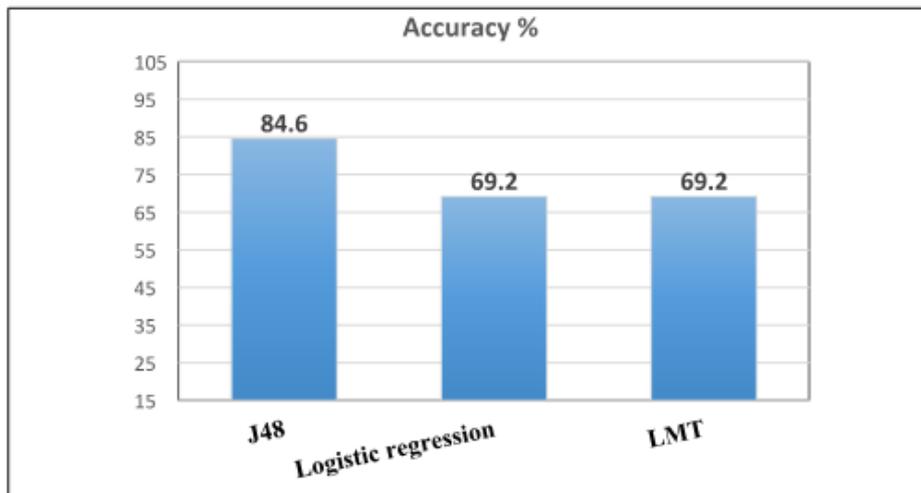

Fig (7): Accuracy% in Peg-IFN monotherapy patients showed J48 optimal accuracy at 84.6%. While LMT was 69.2%.
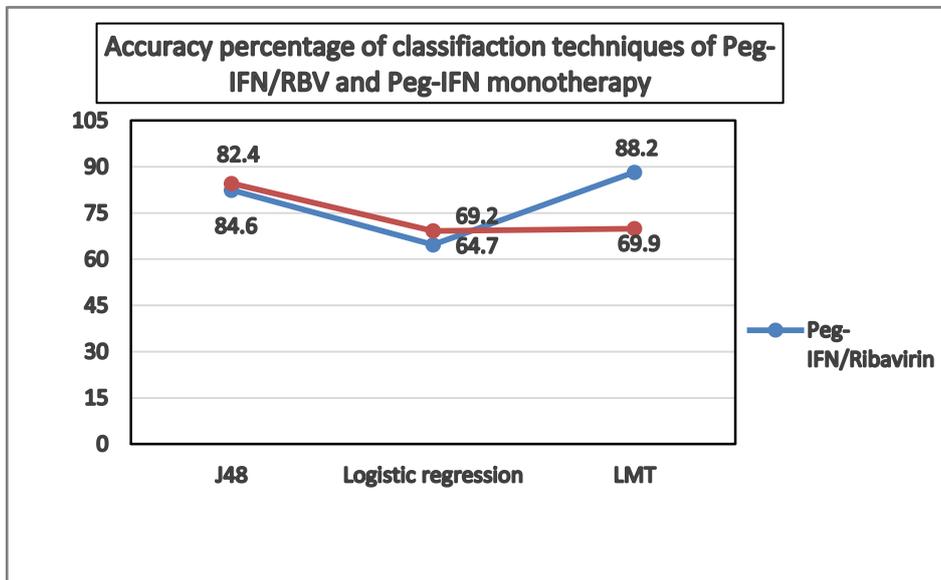

Fig (8). The accuracy percentage of HCV treatment response and classification techniques
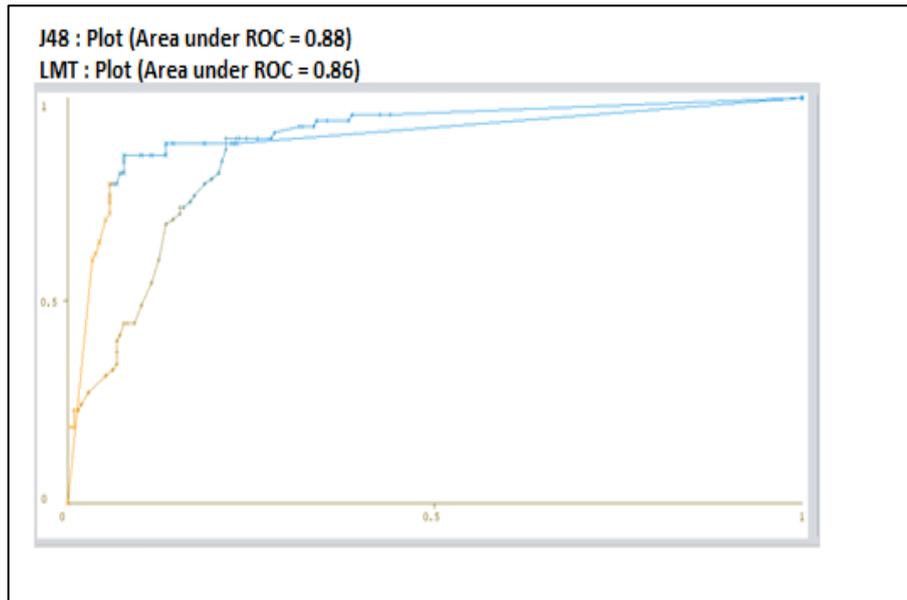
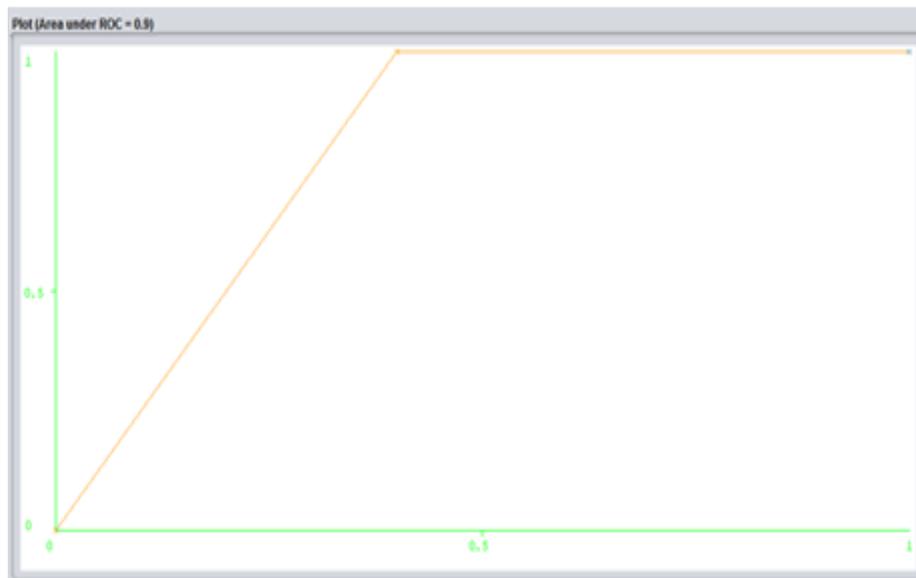**Fig (9): ROC curves of J48 and LMT classifiers for (Peg-IFN/Ribavirin) treatment**

Fig (10): ROC curves of J48 classifier for (Peg-IFN) monotherapy treatment

V. DISCUSION

The main aim of this paper is to predict response of treatment using WEKA data mining tool. We have used algorithms as J48 and logistic regression for our experimentation. These algorithms compare classifier accuracy to each other on the basis of correctly classified instances, mean absolute error and ROC Area. Maximum ROC Area means excellent predictions performance as compared to other algorithms.

We found that how pruning perform well and can influence the accuracy of classification and complexity of tree structure. An interesting question for future research is to proof that datasets also fluent the performance of algorithms. Overall, J48 produces high accuracy of classification and avoid high percentage of

misclassifications. What we need is the simple tree with high accuracy. If the algorithm produces simple tree and low accuracy, it means that algorithm compute high misclassification during instances learning process.

Initial limited trials with interferon plus ribavirin in patients with thalassemia resulted in a 30% increase of blood requirement and prompted an associated increase of chelation therapy.[38] This led to specific contraindication to the use of ribavirin in thalassemia and other hemolytic anemias. Moreover, transfusion requirements increased, as expected, by 56% in combined vs 4.4% in monotherapy patients.

The results of this study demonstrate that rapid clearance of HCV RNA from the serum of patients infected with HCV-GT4 significantly increased the probability of a SVR after 24 weeks of treatment with Peg-IFN/ribavirin. Early clearance of HCV RNA not only increased the likelihood of an end-of-treatment response, but greatly reduced the likelihood of virological relapse during follow-up. The comparison of SVR rates in patients treated for 24 or 48 weeks demonstrates that there was no significant difference between SVR rates among patients who achieved a RVR at week 4.

Our logistic regression model demonstrates that HCV RNA level is the independent significant baseline predictor of RVR. Patients with higher baseline HCV RNA levels were less likely to achieve an RVR, although those with a RVR (9.5% of those with baseline HCV RNA >600,000 IU/mL) were as likely to achieve a SVR as patients with low baseline viral loads 90.5%.

Decision-trees analysis classified Peg-IFN/Ribavirin patients with high probabilities for SVR according to Ferritin baseline levels (P=0.03). Using this model, an estimate of SVR can be rapidly obtained before treatment, and thus may facilitate clinical decision making. Nonetheless, the growing recognition of Ferritin levels as a predictor of response, given its uniqueness and the final decision should be made on a case-by-case basis.

This prospective study showed that low dose ribavirin in chronically HCV infected thalassemia patients is safe, tolerable and effective. Our primary efficacy analysis showed that SVR rate was significantly higher in group (A) patients who received combination therapy with low dose ribavirin in comparison with patients of group (B) who received monotherapy. In addition, HCV-PCR baseline, Ferritin and age were predictors of SVR rate in the total studied population. These findings are consistent with the results of previous studies [12][24]. Moreover, by close observation and frequent transfusion, the rate of severe hemoglobin drop can be minimized. As presented in Table 1 most of major predictors of virological response including HCV RNA level, sex, HCV-PCR, age and liver histological findings appeared to be similar in both groups [24]. However, someone can argue that lack of randomization might underestimate the efficacy of Peg-IFN monotherapy or overestimate the safety of ribavirin. The reason is that patients with more severe thalassemia might have denied ribavirin treatment. This is unlikely in our opinion because comparison of baseline serum ferritin level shows that those who accepted ribavirin therapy had a significant elevated serum ferritin level and subsequently a more severe thalassemia than those who declined to receive ribavirin [25], our study shows 74% ferritin baseline levels in both groups. This analysis indicated that the major impact of low dose ribavirin was in thalassemia major patients with age older than 24 years of age, a low serum ferritin (< 2600 ng/mL), an elevated ALT and liver fibrosis of 0-4 HAI, history of splenectomy and viral load ≤ 600,000 IU/mL. This later observation has been reported by others as well [12].

To evaluate the association between IL28B polymorphisms and the stage of liver fibrosis both combined therapy and monotherapy patients (P=0.59). In TM patients, hepatic iron overload and HCV infection are the major risk factors for progression of fibrosis [23], this study showed (P=0.5) in both treated groups.

VI. CONCLUSION AND FUTURE WORK

Patients with thalassemia major infected with HCV-G4 are the most resistant to treatment and therefore are commonly selected for studies involving new antiviral agents. Our analysis may indicate that therapy with low dose ribavirin may be considered as the treatment of choice in certain patients of thalassemia major patients but not all patients. In clinical practice, the possible applications of logistic regression and decision-tree algorithms; can improve confidence in the results. In the future, more data sets will be used to train other classifiers and to try more experiments. Also other techniques will be applied and more than one technique will be combined to reach as high accuracy as possible.

REFERENCES

[1]. R. Mahmoud, A. El-Mazary and A. Khodeary, "Seroprevalence of Hepatitis C, Hepatitis B, Cytomegalovirus, and Human Immunodeficiency Viruses in Multitransfused Thalassemic Children in Upper Egypt", Advances in Hematology, vol. 2016, pp. 1-7, 2016.

[2]. S. Ansari, M. Ashraf, M. Bohray, S. Erum, M. Ahmed, F. Raza, T. Shamsi, T. Farzana, K. Perveen, I. Ansari and M. Ahmed, "Molecular epidemiology of β-thalassemia in pakistan: Far reaching implications", Indian Journal of Human Genetics, vol. 18, no. 2, p. 193, 2012.

[3]. A.El-Beshlawy, N. Kaddah, A. Moustafa, G. Mouktar, and I. Youssry, "Screening for β-thalassaemia carriers in Egypt: significance of the osmotic fragility test," Eastern Mediterranean Health Journal, vol. 13, no. 4, pp. 780–786, 2007.

[4].   M.G. Ghany, D.B. Strader, D.L. Thomas, L.B. Seeff,"American Association for the Study of Liver Diseases, Diagnosis, management, and treatment of hepatitis C," an update, *Hepatology*.,vol. 49 no. 4, pp. 1335-1374, 2009.

[5].   National Heart Lung Blood Institute site, http://www.nhlbi.nih.gov/health/dci/Diseases/Thalassemia /Thalassemia_WhatIs.html

[6].   Y.J. Sun, J.H. Li, X.Y. Wang, "The diagnostic value about the MCV and red blood cell fragility test in thalassemia screening," Chinese Journal of Birth Health & Heredity, vol. 15, no. 8, pp. 115–116, (2007).

[7].   S. Sandoughdaran, S.M. Alavian, H. Sharafi, B. Behnava, S. Salimi, L. Mehrnoush, P. KarimiElizee, M. Keshvari, "Efficacy of Prolonged Treatment With Pegylated Interferon (Peg-IFN) and Ribavirin in Thalassemic Patients With Hepatitis C Who Relapsed After Previous Peg-IFN-Based Therapy," *Hepat Mon*., vol. 15, no. 1,e23564, Jan 2015.

[8].   K. Falasca, C. Ucciferri, P. Mancino, V. Gorgoretti, E. Pizzigallo, J. Vecchiet,"Use of epoetin beta during combination therapy of infection with hepatitis c virus with ribavirin improves a sustained viral response," *J Med Virol.,*vol. 82, no. 1, pp. 49–56, Jan 2010.

[9].   L. Sharvadze, T. Tsertsvadze, N. Gochitashvili, T. Kakabadze, E. Dolmazashvili, "IFN/Rbv treatment induced anemia and its correction with epoetin alpha in patients with hepatitis C,"*Georgian Med News*.,vol. 137, pp. 62–5, Aug 2006.

[10].   F.A. Ardalan, M.R. Osquei, M.N. Toosi, G. Irvanloo,"Synergic effect of chronic hepatitis C infection and beta thalassemia major with marked hepatic iron overload on liver fibrosis: a retrospective cross-sectional study,"*BMC Gastroenterol*.,vol. 4, pp. 17, Aug 2004.

[11].   K. Soe, Y. Hishikawa, Y. Fukuzawa, N. Win, K.S. Yin, K.M. Win, et al.,"Possible correlation between iron deposition and enhanced proliferating activity in hepatitis C virus-positive hepatocellular carcinoma in Myanmar (Burma),"*J Gastroenterol*.,vol. 42, no. 3, pp. 225–35, 2007.

[12].   S.M. Alavian, S.V. Tabatabaei,"Treatment of chronic hepatitis C in poly transfused thalassaemic patients: a meta-analysis,"*J Viral Hepat.*,vol. 17, no. 4, pp. 236–44, 2010.

[13].   J.C. McHutchison, et al. "Predicting response to initial therapy with interferon plus ribavirin in chronic hepatitis C using serum HCV RNA results during therapy", J Viral Hepat; 8: 414-20, 2001.

[14].   E.M.F. El Houby,"A Framework for Prediction of Response to HCV Therapy Using Different Data Mining Techniques"*Advances in Bioinformatics.*, 11 Dec 2014.

[15].   T. Asselah, G. Esmat, F. Sanai, I. Goulis, D. Messinger, G. Bakalos and I. Waked, "Simple Predictive Model for Identifying Patients with Chronic Hepatitis C and Hepatitis C Virus Genotype 4 Infection with a High Probability of Sustained Virologic Response with Peginterferon Alfa-2a/Ribavirin: Pooled Analysis of Data from Two Large, International Cohort Studies", Advances in Therapy, vol. 33, no. 10, pp. 1797-1813, 2016.

[16].   W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, S. Fucharoen, "Thalassaemia classification by neural networks and genetic programming", Information Sciences, vol. 177, pp.771– 786, (2007).

[17].   E. El-Sebakhy, M. Elshafei, "Thalassemia Screening Using Unconstrained Functional Networks Classifier", In Proc. IEEE International Conference on Signal Processing and Communications (ICSPC 2007), vol. 1, pp. 1027 – 1030, 2007.

[18].   N. Zayed, A. Awad, W. El-Akel, W. Doss, M. El-Raziky, M. Mabrouk,"Low Serum Alpha-fetoprotein Level an Important Predictor for Therapeutic Outcome in Egyptian Patients with Chronic Hepatitis C: A Data-Mining Analysis," *J Gastrointest Dig Syst.,* vol. 4, no. 6, 2014.

[19].   M. Hasani, A. Hanani: "Automated Diagnosis of Iron Deficiency Anemia and Thalassemia by Data Mining Techniques", IJCSNS International Journal of Computer Science and Network Security, vol.17, no.4, 2017.

[20].   P. Harmatz, M. Jonas, J. Kwiatkowski, E. Wright, R. Fischer, E. Vichinsky, P. Giardina, E. Neufeld, J. Porter and N. Olivieri, "Safety and efficacy of pegylated interferon -2a and ribavirin for the treatment of hepatitis C in patients with thalassemia", Haematologica, vol. 93, no. 8, pp. 1247-1251, 2008.

[21].   D. W. Hosmer, S. Lemeshow. Applied Logistic Regression (2nd ed.), 2000, Wiley. ISBN 0-471-35632-8.

[22].   Vijay Kumar Mago and Nitin Bhatia, Cross-Disciplinary Applications of Artificial Intelligence and Pattern Recognition: Advancing Technologies, 2011.

[23].   E. Angelucci and F. Pilo, "Treatment of hepatitis C in patients with thalassemia", Haematologica, vol. 93, no. 8, pp. 1121-1123, 2008.

[24].   H. Al Ashgar, A. Helmy, M.Q. Khan, K. Al Kahtani, M. Al Quaiz, M. Rezeig, et al., "Predictors of sustained virological response to a 48-week course of pegylated interferon alfa-2a and ribavirin in patients infected with hepatitis C virus genotype 4", Ann Saudi Med., 29(1):4–14, 2009.