Classification and Comparison of Hepatitis-C using Data Mining Technique

[1]Saddam Hussain Malik, [2]Dr. Husnain Mansoor Ali

[1,2]Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi Pakistan

[1]saddamhussainmalik1987@gmail.com
[2]husnain.mansoor@szabist.edu.pk

Abstract— The major focus in this paper is to get the factors that shows the significance in predicting the risks of virus of hepatitis-C. 2 datasets were used for this purpose the first one is gathered from UCI Repository and the second one is taken from Zahid Medical Centre with the help of Dr. Abdul Fateh. There are nineteen features and a class feature with classification in binary. The first data set that is gathered from UCI repository contains 155 records with missing values in most of them in order to reduce this technique of normalization is applied. Now for qualitative approaches for data reduction as well as quantitative the binary logistic regression is used. The first result gathered from the Zahid Medical Centre gave us 58% accuracy result using these techniques. And second result using these procedures produced about 90% accurate classification. Our approach gives good classification rate only by using total 37% fields.

***Keywords*—**Data Mining, Regression, Logistic Regression, Normalization, Hepatitis-C, Principal Component Analysis.

I. INTRODUCTION

Hepatitis is actually the burning of the cells of the liver and it can also be said that it is basically the swelling of the liver. There are different forms of the hepatitis viruses, on the basis of the type of the infections of the virus which are mainly A, B and C. At worldwide or global stage about 130–150 million persons predicted which were affected by the virus of hepatitis type C [1].

Not only nowadays but Hepatitis has always been a major issue regarding health. So there is a high probability or higher chances that you will encounter the people associated with this virus. So many of the people who have hepatitis virus don't show that they are infected by this virus, they don't let other people know that they are infected with the hepatitis virus. So because of this there are a lot of chances that people around these infected people also get infected with the virus of hepatitis. That's because they become careless rather than being careful that they should avoid doing things because of which other people can get infected. Though there are only few precautious steps that they should follow in order to avoid the spread of virus to other people.

Now in our country Pakistan the doctors use different methods for the Hepatitis-C two of them are PCR and Elisa method. PCR method is basically used when the patient undergoes antibodies test if the result for the antibodies is positive then the PCR test is used by the doctors. The use of PCR test is to find out whether there is a virus living in the patient's body or not. If the virus is not found in the body of the patient it means that patient is not carrier of HCV and it will also state that patient was a carrier of HCV but it ejected its body recently. If the Virus found then patient undergoes liver biopsy and ultrasound. Elisa method is used to measure the antibodies in the patient. An enzyme is used to check the measure of the antibodies which are related to certain infectious condition.

Now for the prediction in medical science is dependent on the study, knowledge and experience of the physicians. Before the prediction of any medical data set it is compulsory for you to at least learn the data set of the disease from the physician in order to go along with the physicians' idea once that knowledge is gained then many people use different algorithms and technique to measure the prediction of the disease from the population samples. For that many people use models, pattern recognition system for the prediction of the pattern of the data sets on the basis of which future values can be predicted and the results in aspect of the accuracy with the help of the tools and technique they achieved is given. So our work is also same we will show you how we predicted the results in our methodology.

II. OBJECTIVE OF THE STUDY

The objective is to predict the results for the Hepatitis-C patients with the help of the datasets. For that firstly one of our data requires the normalization because it has many missing values. After the normalization, data will go under

Principal component analysis. PCA's purpose is to reduce the dimensions for the datasets. We are doing this because our data set has many features. There are 19 features with one output feature that will tell whether the patient having symptoms (19 features) is dead or still alive. Now for our study and research work because the difference in the value type exists like few values are the number representing 'yes' and 'no' our output field class tells us whether the patient is dead or alive represented in number again with 1 as dead and 2 as alive so the metric difference in other values exists like the enzyme's value are in a format like 0.7 or 2.3 so they represent different scale. PCA will be applied in order to check that which features with what scale are important for the data prediction. Different scales can't be used at the same time in the PCA otherwise its mechanism will prefer only one scale type values. Now these dimensions show that how much their importance will influence on our study these dimension will be used for the regression model that we will be using for the forecasting of the future values, with the help of this the classification table will be created which will show you our prediction result with the original result. Linear regression is the most essential and generally utilized prescient investigation. Regression assessments are utilized to portray information and to clarify the relationship between one ward variable and at least one free variables. At the focal point of the relapse investigation is the assignment of fitting a solitary line through a diffuse plot. Regression Logistic is the proper regression investigation to lead when the needy variable is parallel. Like all regression investigations, the calculated regression is a prescient examination. Logistic regression is utilized to portray information and to clarify the relationship between one ward paired variable and at least one metric which could be interim or proportion scale autonomous factors.

III. LITERATURE REVIEW

In relation to our research study there are many researchers who have worked on the problem which regards Hepatitis-C which help doctors and physician a lot in their work field, some of the researchers work is given below.

For the inflammation rate in the liver Ishak system is used made by Ishaq *et al*. [2] in 1995, the inflammation gathered from the observer on the light microscope assigns the value between 0-18. Similarly the scarring factor is rated between 0-6.

Kedziora *et al*. [3] used his approach in which difference between HCV organisms in the present population using trees of Phylogenetics and Hamming distances after which the patients responded negatively and positively when applied therapy was applied.

In 2006 that Hodgson *et al*. [4] used this approach for the inflammatory cells they made an automated system for the infected liver biopsy in hepatitis-C the cell's quantification was carried out in his approach. After the features extraction from biopsy image.

In 2006 Polat *et al*. [5] used function selection for the system of hybrid and fuzzy aid allocation mechanism method for synthetic immune popularity machine which is used for the analysis of the virus of the hepatitis. Resultant accuracy for the class that was obtained by them become 92.59%. The specificity values and sensitivity values for effects from the test consequences were 85% and a 100%.

In 2004 Guan *et al*. [6] used the application of neural network in incidence for forecasting of hepatitis a, they used the arima model and amassed the results after that the outcomes have been compared with the results the Ann model. So their studies concluded that the artificial neural network version is manner higher than the alternative model inside the light of forecasting the hepatitis a incidence that's as it has regression phenomenon.

In 2002 Avendano *et al*. [7] used the population data which are the people having, infected liver cell, T cells, HCV and people with uninfected liver cells with the help of this data he performed forecasting from the earlier dataset, the discussion was then carried out in which the efficiency of the therapy method for the HCV was discussed afterwards they created the model for HCV dynamics.

The administration in patients with perpetual viral hepatitis C and B rely upon sum along with movement in fibrosis of liver in hazard for cirrhosis. For biopsy of liver, customarily thought in the standard reference for arranging liver fibrosis, it's been tried many times in earlier decade by the headway of non-invasive methodologies [8]. These strategies rely on upon specific yet comparing philosophies: a biologic approach, which assesses serum levels of biomarkers of fibrosis, and a physical approach, which measures liver solidness by ultrasound or appealing resonation elastography. Non-invasive procedures were at initially considered and affirmed in patients with unending hepatitis C however are right now used continuously for patients with hepatitis B, diminishing the prerequisite for liver biopsy examination.

Their overview at central focuses and confinements of the non-invasive techniques used to regulate patients with interminable viral hepatitis B or C pollution this approach was given by Castera [9] in the year 2012.

A programmed conclusion framework in light of Neural is organized for hepatitis infection is presented. Jilani *et al*.

[10] programmed conclusion framework manages the blend of highlight extraction also, arrangement. The framework has two phases, which are highlight extraction, diminishment and characterization stages and was presented by them in the year 2011.

IV. METHODOLOGY

Now to get the information yet unknown along with the precious examples from record is sometime alluded with mining of the records. Such phrases statistics global magazine of computer disclosure, records healing, deductive getting to know and exploratory statistics research can be applied as part of surrounding or area of interest of the data mining. To finish special undertakings, a huge variety of calculations are protected in mining of data. Commonly extensions for the mining of the data is shared out with prescient along with expressive degrees which are associated to application precise modifications regarding necessary needs for issues. Creating expectation approximately records beforehand recognized outcomes with certain statistics completed through prescient version wherein recognizable proof of examples in information is made by way of elucidating model. For the methods that we are going to use in the study SPSS will be used.

A. Principal Component Analysis:

For the reduction of the dimensions of a vast information is decreased through utilizing essential segment examination which is considered as a standout amongst the most famous and helpful factual strategies [11]. This strategy changes the first information into new measurements. The new factors are framed by taking direct mixes of the first factors it gives us the stacking parameters. The new tomahawks are balanced to such an extent that they are orthogonal to each other with most extreme data pick up. The main essential segment having the biggest difference. Because the instantaneous calculation of lattice b is impractical in highlight trade, the initial step is to determine the covariance network. The subsequent stride through which we can get the values for Eigen and that of covariance network. At ultimate, an immediate exchange is characterized by means of the vector which are denoted by n for Eigen relate with the other values of Eigen which are gathered from the dimension m and the other dimension n area. Relevant tomahawks, which can also be said as the vectors of the Eigen, relate to the values of the Eigen. For maximum part, the preliminary few crucial segments include the general public of the statistics. Utilizing the extent of the evaluation of variations can propose us of how many essential segments to be held from the records set. With the example we will show you how it is used for the selection of the best feature selection.

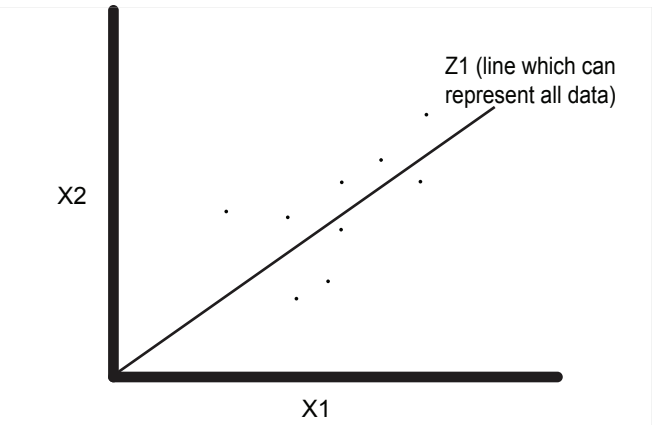

Fig. (1). Feature relationship representation

In figure 1, Z1 represents the line on which we can draw accordingly the data set where all the dataset are close to each other X1 and X2 are basically the features in the data set.

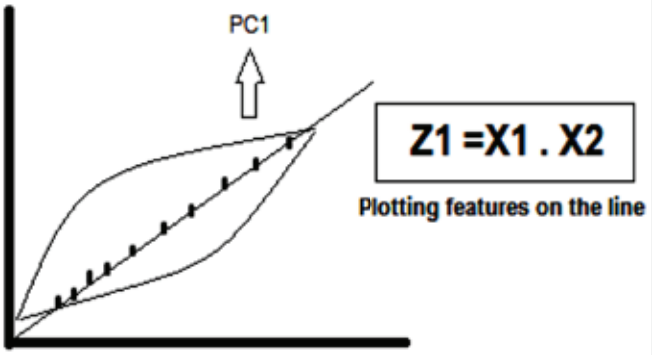

Fig. (2). Feature factor extraction

Now the first component is created in the direction of the line where all the features are put on the line in the direction of the line now the second component will be created in vertical direction and so on as mentioned in figure 2.

Now the question arises that how can this line be achieved or how do we get these components. For that covariance matrix is used, its formula is;

$$Covariance\ Matrix = \frac{1}{n}\sum (xm - \mu).(x - \mu)^T \qquad (1)$$

Where x in the formula is the feature factors in the data μ is the average of the factors and *T* is the transpose. With this the matrix will be achieved.

After this Eigen vectors and values along with corresponding vectors will be achieved. Basically these Eigen help us understand the significance of the feature in the data set and show us through the graph. So for the calculation of the Eigen we follow following formula;

$Ax = \lambda x$

Now Eigen vector corresponding to the largest Eigen value is 1st PC then 2nd PC and so on. For the λ you will get λ1 and λ2 if the two components are selected. Now after all the calculations you will get a value for example let's say the value is 1.5. So this value in the graph will be creating the line from point zero till the end of the feature record. In the same way same thing will be done for second component, that value will be created for it and how close the rest of the records are with the line the value created with the equation are is shown. So that actually implies the relation of the one record with the other is. Now through the application it can be shown as;

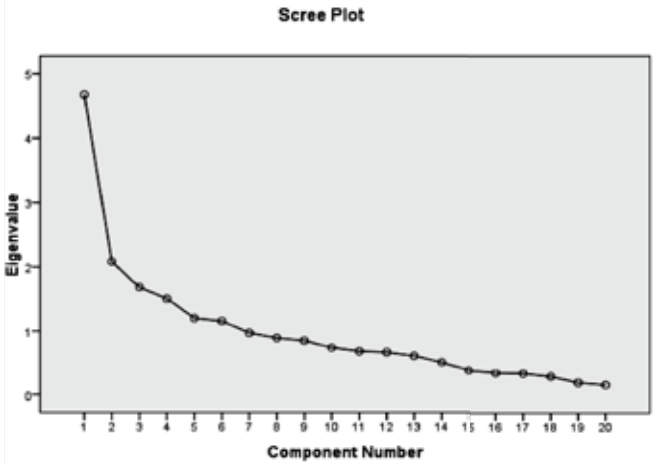

Fig. (3). Features importance with respect to Eigen values

Figure 3 shows the scree plot of the data set with the degree of the importance of the components of the data in aspect of Eigen value which is on y-axis. Along with this plot you will also get other outputs in which the extraction of the component is shown by the application with the variance percentage between the extracted components and the rest of the components of data set. Similarly another table namely component matrix will be provided when applied PC, it will show the linear combination of the PCs with the rest of the records and show the probability of those records with PCs.

B. Logistic and Linear Regression:

The displaying of the likelihood model for occasion happens in aspect with an immediate arrangement indicators the points that are applied also can be said as variables are alluded with regression of the strategic. In which, reaction factors are utilized and in addition not unusual logarithms base and coefficients and signs in my opinion. We can say that achieving logistic regression requires description of analysis, analysis of univariable, co-linearity testing, and analysis of the multivariable and diagnosis of the model.

For comparing decency assault version, this is to be done in model record measurable, and check whether it fits perfectly, through integrity suit data is clarified. Through residual investigation greater part to test decency in assault that is in version gets finalized or finished. Despite the fact that, for a parallel (0-1) result variable, this approach isn't always high-quality. The opportunity ability is a parameters potential which communicates the watched fact's chance. The log-opportunity ability is likewise utilized. Wherein result and the anticipated information likelihood is checked one by one of occasion happening.

So most people ask what is the difference between both linear and logistic, linear is utilized when the wanted yield is required to take a persistent esteem in view of whatever information set is given to the calculation. Assume you need to make a program which would foresee the normal temperature of say tomorrow, in light of specific elements, similar to normal temperature, least temperature, most extreme temperature, and so forth of past week. Since this issue needs yield as an estimation of ceaseless nature, it is named a direct regression issue.

Logistic assume your issue was to not yield the temperature, but rather the sort of climate that tomorrow may have for instance sunny, shady, stormy, blustery. This issue will give a yield having a place with a specific arrangement of qualities predefined, subsequently it is essentially grouping your yield into classes. Arranging issues can be either double 0/1 No/Yes or multi class like the issue portrayed previously. Calculated regression is utilized as a part of ordering issues of machine learning.

Now for linear regression there are only 3 steps required to achieve it, which are correlation, estimation of the model whether the model is fitting in the line, and last is the evaluation of the model worthiness. Now in the process of linear regression the terms R and other term R squared is used. So the purpose of R is basically to check the correlation between the features i.e. suppose there are two features R would tell you that if one is increasing the other is increasing or decreasing in the value, this is strong positive/negative correlation and if no effect in increasing or decreasing in fellow feature occurs than it means that there is no correlation between them. So R can be stated as;

$$R = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \quad (2)$$

(2)

$$R^2 = (R)^2$$

$$\int xy = Sxy/Sx\,Sy \quad (3)$$

Where S represents standard deviation which is;

$$S = \sqrt{\sum(x-\bar{x})^2/n-1} \quad (4)$$

And variance has two type, one is taken of population and other is taken from population also called sample variance;

$$\sigma^2 = \sum(x_i-\bar{x})^2/N \text{ (For population variance)} \quad (5)$$
$$s^2 = \sum(x_i-\bar{x})^2/n-1 \text{ (For population sample variance)} \quad (6)$$

So we will get many of the tables discussing variance for that these formulas will be useful. Variance is useful it actually tells us about the range as well as spread of the data. You can also plot the graphs using the Spss, other than that there is another table created in the linear regression output it is called ANOVA, which basically tells us the reaction time of the features.

So after getting the PCs we will save those PCs in the different columns and now only these components will be used for the prediction of the single record which is dependent on the predictors.

C. Dataset:

There are datasets one of the dataset is taken from uci repository it consists of 19 fields with one output area. The result subject has values which shows whether the affected person is lifeless or alive. For this purpose to record checking for the virus existence of hepatitis is carried out.

The primary hepatitis dataset carries a hundred and fifty five statistics with missing values. The second one dataset carries one hundred statistics of the patient accrued from the Zahid clinical centre with the assist of Dr. Abdul Fateh this facts does now not have any lacking values.

Following are the values of the datasets that I have gathered from UCI and The Medical Centre. In the dataset the class shows the status of live or death of the patient denoted by 1 and 2. Age shows the age patient has. Sex shows the gender of the patient also denoted by 1 and 2. Same is the case with antiviral, malaise, steroid, anorexia, fatigue, spiders, ascites, liver firm, varices, liver big, spleen palpable and histology are denoted with 1 and 2 that whether the patient has these symptoms are not. The remaining values in the dataset are the enzymes which have their values as mentioned in table 1.

Table 1. Data set fields

Age (of patient)
Class (Patient alive or Patient dead)
Antiviral
Sex (Male or Female)
Liver Firm (liver hardness)
Malaise (feeling of discomfort)
Fatigue (feeling of tiredness)
Liver Big (liver inflammation)
Steroid
Spleen Palpable (checking liver hardness)
Anorexia (lack of appetite for food)
Spiders (prominent veins)
Bilirubin (enzyme measure)
Varices (tissue decomposition)
Albumin (enzyme measure)
SGOT (enzyme measure)
Alk Phosphate (enzyme measure)
Histology
Protime (enzyme measure)
Ascites (Water presence)

V. RESULTS

Following are the results that we gathered using above methodology for the prediction of the data that we gathered from Zahid Clinical centre.

Firstly we went through important components through which factors were created, following are the factors that were extracted from the data in table 2.

Table 2. Combination of components extracted with features

	Component					
	1	2	3	4	5	6
Steroid	.084	.260	.001	.411	-.145	.538
Age	-.289	.229	.323	-.505	.078	.258
Fatigue	.541	.393	.331	.154	-.004	-.160
Sex	.038	-.293	.157	.122	.266	.598
Liver Big	.005	.564	-.477	.269	.294	.188
Spiders	.651	.094	-.136	-.092	.245	-.349
Malaise	.512	.429	.544	.195	.047	-.029
Alk Phosphate	-.577	-.160	.276	.319	.157	-.106
Antivirals	-.332	.298	-.008	.434	-.345	-.257
Albumin	.705	-.269	.027	.169	-.216	.111
Spleen Palpable	.408	.041	-.382	-.150	.423	-.077
Histology	-.607	.219	.199	-.220	-.210	-.096

Extraction Method: Principal Component Analysis.

a. 6 components extracted.

Table 3. Features extracted

1) Antiviral
2) Steroid
3) Malaise
4) Fatigue
5) Liver Big
6) Anorexia

Table 3 shows the components extracted from the implementation of data set on principal component with the scree plot given below:

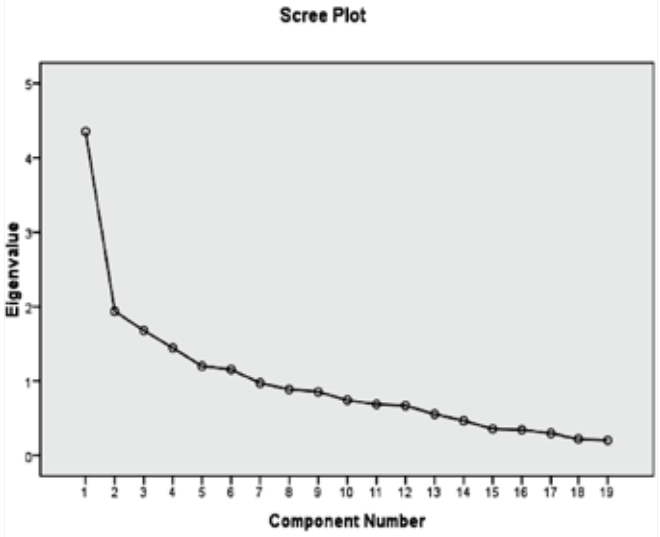

Fig. (4). Feature importance demonstration with scree plot

The scree plot in figure 4 provides the Eigen values with respect to the values given in the data set. The first six values are the important factors whose covariance with the rest of the values in the record is effective. Now the components factors created will be 6 as result given in the component matrix table 4.

Table 4. Creation of 6 principal components

Anorexia_f	Numeric	11	5	None	None	13	Right	Scale
Steroid_f	Numeric	11	5	None	None	13	Right	Scale
Malaise_f	Numeric	11	5	None	None	13	Right	Scale
Sex_f	Numeric	11	5	None	None	13	Right	Scale
Antiviral_f	Numeric	11	5	None	None	13	Right	Scale
Fatigue_f	Numeric	11	5	None	None	13	Right	Scale

Table 5. Values generated through linear regression

	B	Std. Error	Beta	T	Sig.
Anorexia_f	.191	.033	.514	5.870	.045
Steroid_f	-.120	.033	-.324	-3.697	.000
Malaise_f	.026	.033	.070	.797	.426
Sex_f	.092	.033	.248	2.830	.006
Antiviral_f	-.015	.033	-.041	-.468	.641
Fatigue_f	.016	.033	.043	.491	.625

a. Dependent Variable: Class

Above table 5 was generated using linear regression.

So after applying the results gathered on the logistic regression we got the classification table in which our model prediction result for dead and alive patients is shown.

Table 6. Prediction result of Zahid medical centre

		Predicted		
		Die	Live	% correct
Observed	Die	16	15	51.61%
	Live	27	42	60.86%
	Overall %			58%

By using our methodology we are not getting much accurate result, it shows the accuracy of 58% in table 6. Actual result was 31 deaths predicted result is 16 deaths there is a lot of difference between the actual and predicted death, moving on actual live patient are 69 and predicted are 42 still we are close though but it would be more effective if we took more than one sample and apply or technique on it.

Now applying same technique on the 2nd data set the result gathered are;

Table 7. Prediction result of UCI Data

		Predicted		
		Die	Live	% correct
Observed	Die	8	5	61.54%
	Live	3	64	95.52%
	Overall %			90%

So when our technique applied on the UCI data we gathered some good result, giving us 90% accuracy result in table 7. Which is good thing, so total deaths were 13 our predicted deaths are 8 and live patients were observed to be 67 but predicted are 64.

So our approach proves out to be positive on this data set. After applying the principal component on the UCI data we got antiviral, malaise (uneasiness, discomfort, illness), liver firm (hardness of liver), steroid, fatigue (tiredness), and liver big (inflammation in liver) variables. Now they proved out to be more significant when it came to correlation of them between the rest of the features using the linear regression and after applying logistic we got above result.

VI. CONCLUSION & FUTURE WORK

The discussion in the paper is which features are to be selected for consideration, with the help of which we get more accurate result when it comes to the prediction. So for that we used the reduction technique which narrows down the feature

Which should be used for prediction. Afterwards the linear and logistic were used to check the significance of the features and finally give the classification accuracy. But we observed that for different samples we get different result like it was observed between two different regions data.

For the future studies we need to at least get more sample data from one population space so that we can pin point the dimensions. Though we can't fully rely on the application result but we can consider only those features which are going to be more important for prediction result for that population sample. Other than that we need to focus on the rest of the features specially the enzymes, for the feature selection because they are also important symptoms which we should not ignore when it comes to the disease related to different virus of hepatitis. For these we need to apply different techniques and compare them with the result that we gathered and decide the modifications which need to be done in our techniques in future.

REFERENCES

[1] *WHO, Hepatitis C, Fact Sheet No. 164* [Online]. Available:http://www.who.int/mediacentre/factsheets/fs164_apr2014/en/

[2] K. Ishak, A. Baptista, L. Bianchic, F. Callead, J. D. Groote, F. Gudat, H. Denk, V. Desmet, G. Korb, R. N. M. MacSween, M. J. Phillips, B. G. Portmann, H. Poulsenm, P. J. Scheuer, M. Schmid, and H. Thalero, "Histological grading and staging of chronic hepatitis," *Journal of Hepatology,* vol. 22, no. 6, pp: 696–699, 1995.
DOI: 10.1016/0168-8278(95)80226-6

[3] P. Kedziora, M. Figlerowicz, P. Formanowicz, M. Alejska, P. Jackowiak, N. Malinowska, A. Fratczak, J. Blazewicz and M. Figlerowicz, "Computational Methods in Diagnostics of Chronic Hepatitis C", *Bulletin of the Polish Academy of Sciences-Technical Sciences*, vol. 53, no. 3, pp: 273-281, 2005.

[4] S. Hodgson, R. F. Harrison and S. S. Cross, "An automated pattern recognition system for the quantification of inflammatory cells in hepatitis-C-infected liver biopsies," *Image and Vision Computing*, vol. 24, no. 9, pp: 1025–1038, 2006.
DOI: 10.1016/j.imavis.2006.02.019

[5] K. Polat and S. Gunes, "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation", *Digital Signal Processing*, vol. 16, no. 6, pp: 889–901, 2006.
DOI: 10.1016/j.dsp.2006.07.005

[6] P. Guan, De-S. Huang and B-S. Zhou, "Forecasting model for the incidence of hepatitis A based on artificial neural network," *World Journal of Gastroenterology*, vol. 10, no. 24, pp: 3579–3582, 2004.
DOI: 10.3748/wjg.v10.i24.3579

[7] R. Avendano, L. Esteva, J. A. Flores, J. L. F. Allen, G. Gómez and J. López-Estrada, "A Mathematical Model for the Dynamics of Hepatitis C," *Journal of Theoretical Medicine*, vol. 4, no. 2, pp: 109-118, 2002.
DOI: 10.1080/10273660290003777

[8] I. A. Moneim and G. A. Mosa, "Modeling the Hepatitis C with Different Types of Virus Genome", *Computational and Mathematical Methods in Medicine*, vol. 7, no. 1, pp: 3-13, 2006.
DOI: 10.1080/10273660600914121

[9] L. Castera, "Noninvasive methods to assess liver disease in patients with hepatitis B or C." *Gastroenterology*, vol. 142, no. 6, pp: 1293-1302, 2012.
DOI: 10.1053/j.gastro.2012.02.017

[10] T.A. Jilani, H. Yasin, and M. M. Yasin, "PCA-ANN for classification of Hepatitis-C patients." *International Journal of Computer Applications*, vol. 14, no. 7, 2011.

[11] K. Polat, and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing*, vol. 17, no. 4, pp: 702–710, 2007.
DOI: 10.1016/j.dsp.2006.09.005