Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach*

Heba Ayeldeen*1, 2, Olfat Shaker [1,3], Ghada Ayeldeen [1,3], Khaled M. Anwar [2]
1 ISI Research Lab - www.isirlab.net
2 Faculty of Computers and Information, Cairo University, Egypt
3 Department of Medical Biochemistry and Molecular Biology, Faculty of Medicine, Cairo University, Egypt

***Abstract*—Using Information systems and strategic tools for medical domains is constantly growing. Automated medical models play an important role in medical decision-making, helping physicians to provide a fast and accurate diagnosis or even prediction. Making use of the knowledge or even in the early stages of knowledge acquisition, different statistical mining and machine learning tools can be used. For instance predicting whether the patient with Hepatitis C virus has also liver fibrosis or not is one of the concerns. In case the prediction result is true, in what stage is the fibrosis. To easily reach to this knowledge without costly diagnostic routine laboratory tests there should be a fully integrated system. Therefore in this study we used machine learning technique model based on decision tree classifier to predict individuals' liver fibrosis degree. Results showed that by using decision tree classifier accuracy is 93.7% which is higher range than what is reported in current researches with similar conditions.**

Keywords*—Hepatitis C virus, Liver fibrosis, Decision Tree classification, Machine Learning techniques.

I. Introduction

Of the most common causes of acute liver disease are viral hepatitis (particularly hepatitis A, B and C) [1]. Taking sample of the liver (Liver biopsy), remains the current main standard for assessment and diagnosis of fibrosis and inflammation of the liver. This process is very costly and needs a lot of time, resources and effort as well [1], [2]. Therefore, we aim to identify the most informative biomarkers/variables provided from individual laboratory tests through involving the highly statistical data mining techniques (i.e. statistical machine learning techniques) to predict the hepatic fibrosis in Egyptian patients infected with hepatitis C.

Egypt has the highest prevalence of HCV in the world reaching 14.7% of the population. Estimated 11 millions anti-HCV-positive persons. HCV is a major cause of chronic liver diseases and liver cirrhosis (which is the late stage of fibrosis/inflammation of the liver) [3].

A. Serum biomarkers

There are routine functions for HCV: aspartateaminotransferase (AST) enzyme, alanineaminotransferase (ALT) enzyme, serum albumin (ALB), total bilirubin (T.BIL) and direct bilirubin (D.BIL). These are the blood tests that are commonly used for the assessment of liver disease, measuring levels of serum ALT and AST; alkaline phosphatase; direct and total serum bilirubin; and albumin [1], [4], [5].

Further testing can be needed that is used to detect the existance of fibrosis in patients with HCV rather than the biopsy. Of these biomarkers are Gamma-Glutamyl Transpeptidase (GGT), Hyaluronic Acid (HA), α 2-macroglobulin (α 2-MC) and apolipoprotine A2 (ApoA2) and others [1], [6].

Each of these biomarkers detect the degree of fibrosis. These serum markers, have the potential to replace liver biopsy. The interpretation of liver fibrosis stages is mainly classified into five different stages which are [3], [6]:

- **F0** =no fibrosis
- **F1** =portal fibrosis without septa
- **F2** =fibrosis with few septa
- **F3** =fibrosis with numerous septa
- **F4** =cirrhosis

Numerous studies have shown that machine learning and data mining are powerful tools in the medical sector with great diagnostic potential of diseases due to its ability to discover the secured hidden predictive patterns from medical databases [7-10]. Therefore, we aim to early identify the stage of fibrosis in hepatic C Egyptian patients; provided from individual laboratory tests and selecting the significant parameters. We followed the decision tree approach as a machine learning techniques. We designed a machine learning model that consists of six phases that are described in more details the following sections.

The rest of the paper is organized as follows: Section II presents the proposed Machine Learning model for classification and early predict of the fibrosis degree. While section III describes the sample population dataset used in the case study as well as the variables identification. Section IV briefly describes how the data is stored and well indexed. Section V shows mathematically how the features/variables can be selected according to the level of significance with respect to the problem domain. Section VI demonstrates how data can be classified by using decision tree showing the input and the output variables. Section VII shows how knowledge can be extracted and be used as prediction. Last but not least is section VIII which states the medical interpretation after using decision tree classifier on the hepatic fibrosis data we have. Finally the conclusion and future work is illustrated in section IX.

*All the hepatic fibrosis patients involved in the study were diagnosed at department of Biochemistry and Molecular Biology of Kasr Alainy Hospital of Cairo University. For all participants in this study, written informed consent was obtained as delineated by the protocol which was approved by the Ethical Committee of Cairo University.

978-1-4673-9669-1/15/$31.00

II. Machine Learning model

What is ”machine learning” anyway? This is a question that is asked everyday. Simply machine learning is everything, the field is quite vast and growing. Machine learning (ML) is a subfield of computer science that involved plenty of subfields [9], [11].

”Field of study that gives computers the ability to learn without being explicitly programmed”- [12].

This is how Arthur Samuel defined machine learning back in the 50's. Arthur Samuel was an American pioneer in the field of computer gaming, artificial intelligence, and machine learning [12].

Machine learning can be related to other fields like data mining, statistics, Knowledge discovery and others. According to the problem domain, technique of ML is selected and wisely used to achieve the required objective [9].

In order to achieve the preceding objective of predicting the liver fibrosis degree we carried out the framework shown in Fig. 1. In the process of data collection and preparation, data is collected from the department of Medical Biochemistry and Molecular Biology, Faculty of Medicine, Cairo University. The data includes laboratory tests and fibrosis markers, the outlier data is removed. The whole data is divided into two independent sets: training set (learning) and testing set (classification). The next module is Data storage where MYSQL 5.2 Community Server 5.5.27 is used to store the data. In the feature selection module we find the appropriate features for representing the input patterns using graphical and statistical feature selection strategies. In the classification module we used the training data as the input for the classifier of the features selected. More-or-less decision tree classifier is developed to analyze the data and predict individual's liver fibrosis degree. Finally, we evaluate the model results from medical perspective in the Medical Interpretation module to make use of the knowledge retrieved for prediction.

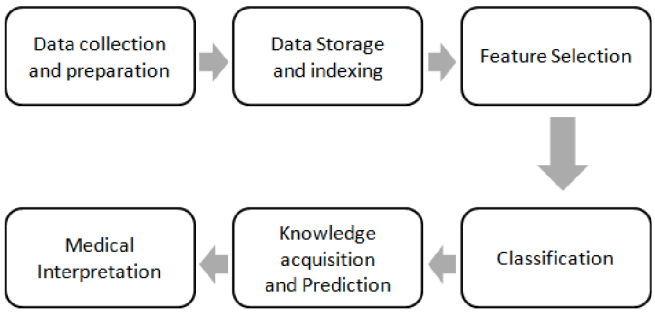

Fig. 1: Framework of machine learning model to predict the individual stages of hepatic fibrosis in patients with HCV.

III. Data collection and preparation

The objective of this phase is to prepare and list the main parameters that directly affect the liver fibrosis and its degree and to define the upper and lower criteria. After the data is collected, the routine liver function tests as well as serum functions are well organized and represented.

A. Variable Identification

In this step all the variables or effective parameters are well identified. There are ten variables used in this study. The age parameter is considered as one of the variables although it has been proven that it is not statistically significant. The other nine parameters are classified into two classes (Routine function tests and Serum tests). The routine function tests variables are: aspartateaminotransferase (AST) enzyme, alanineaminotransferase (ALT) enzyme, serum albumin (ALB), total bilirubin (T.BIL) and direct bilirubin (D.BIL). While the serum tests are: Gamma-Glutamyl Transpeptidase (GGT), Hyaluronic Acid (HA), α 2-macroglobulin (α 2-MC) and apolipoprotine A2 (ApoA2).

B. Data sampling

Participants in this study were 100 chronic HCV patients from both genders, where age ranges between 19-60 years old. All the hepatic fibrosis patients involved in the study were diagnosed at department of Biochemistry and Molecular Biology of Kasr Alainy Hospital of Cairo University. For all participants in this study, written informed consent was obtained as delineated by the protocol which was approved by the Ethical Committee of Cairo University. Fibroscan machine is used to detect the degree of fibrosis classified into: No Fibrosis (F0), Portal Fibrosis (F1), Few Septa (F2), Many Septa (F3), and Cirrhosis (F4).

IV. Data storage and indexing

This module is concerned with Data storage where MYSQL 5.2 Community Server 5.5.27 is used to store the data. The names for the patients are kept confidential so data is indexed where each record is given an ID that is unique and considered as the primary key of patient. The presentation software PHP 5.5 is used as an interface layer.

The figure below shows sample of the data structured in the database for HCV Egyptian patients.

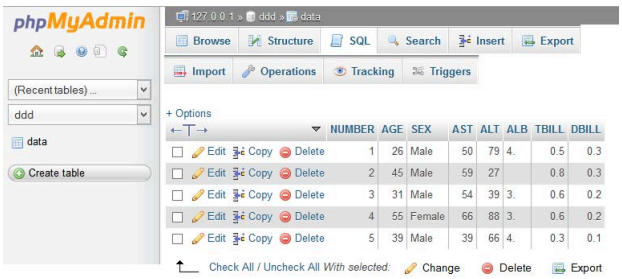

Fig. 2: Sample data for HCV Egyptian patients .

V. Feature selection

A. Data exploration

According to the patient list we have, the dominant patients were males with 67% with a mean age of 38 years. As determined by the Fibroscan and by using 95% confidence interval of the mean for all laboratory tests the inclusion of the data for males were (F0:6%, F1: 29%, F2: 16%, F3: 7%,F4: 9%). Table 1 shows the patients' baseline variables included in this study.

TABLE I: Patients' baseline variables.

Variables	Mean ± SD	Minimum	Maximum
Age	40.29±10.98	19	59
AST(U/L)	52.61±31.61	11	180
ALT(U/L)	55.24±33.0	13	172
ALB(g/L)	4.19±0.49	3	5.2
T.BILL(μ/L)	0.81±0.21	0.3	1.5
D.BILL(μ/L)	0.34±0.14	0.1	0.8
GGT(U/L)	68.30±52.71	11.3	253.4
HA(ng/ml)	0.16±0.13	0.03	1.05
α 2-MC (g/L)	2.59±0.54	1.35	4.09
Apo-A2 (g/L)	0.42±0.20	0.08	0.94

B. Selecting significant variables

According to the Fibroscan machine, there are five different degree of fibrosis identified, F0, F1, F2, F3, F4 and F5. To detect the level of significant of all variables across these degrees we calculated the P-value comparing it to all classes. Table 2 and 3 show description data statistical calculation of all degrees/variables.

TABLE II: ANOVA table of biomarkers and fibrosis degrees.

Biomarkers	Hepatic Fibrosis degree - (Mean ± SD) F0 n=7	F1 n=38	p-value
Age	35.85±10.23	35.4±9.310	0.3633
AST(U/L)	45.14±18.41	43.92±28.14	<0.001*
ALT(U/L)	34.14±23.26	50.71±30.76	<0.001*
ALB(g/L)	3.85±0.69	3.78±0.905	0.0751
T.BILL(μ/L)	0.82±0.22	0.74±0.194	<0.001*
D.BILL(μ/L)	0.32±0.125	0.30±0.135	<0.001*
GGT(U/L)	36.11±28.35	39.17±33.43	<0.001*
HA(ng/ml)	0.10±0.048	0.12±0.077	<0.001*
α 2-MC (g/L)	1.68±0.36	2.23±0.24	<0.001*
Apo-A2 (g/L)	0.81±0.096	0.53±0.169	<0.001*

TABLE III: ANOVA table of biomarkers and fibrosis degrees.(Cont'd)

Biomarkers	Hepatic Fibrosis degree - (Mean ± SD) F2 n=28	F3 n=12	F4 n=15	p-value
Age	40.60±11.80	47.00±10.34	48.73±6.05	0.3633
AST(U/L)	53.85±35.47	56.58±25.07	72.60±35.07	<0.001*
ALT(U/L)	55.03±30.84	60.16±24.12	74.33±45.17	<0.001*
ALB(g/L)	3.82±0.611	3.58±0.514	3.53±0.51	0.0751
T.BILL(μ/L)	0.846±0.211	0.82±0.200	0.89±0.22	<0.001*
D.BILL(μ/L)	0.36±0.147	0.33±0.177	0.393±0.127	<0.001*
GGT(U/L)	88.18±56.43	81.98±28.11	109.09±63.54	<0.001*
HA(ng/ml)	0.13±0.067	0.27±0.258	0.31±0.098	<0.001*
α 2-MC (g/L)	2.75±0.29	3.53±0.30	2.88±0.07	<0.001*
Apo-A2 (g/L)	0.31±0.114	0.27±0.090	0.31±0.123	<0.001*

Note: The P-values in tables 2 and 3 are relative to all fibrosis degrees (F0,F1.F2.F3 and F4).

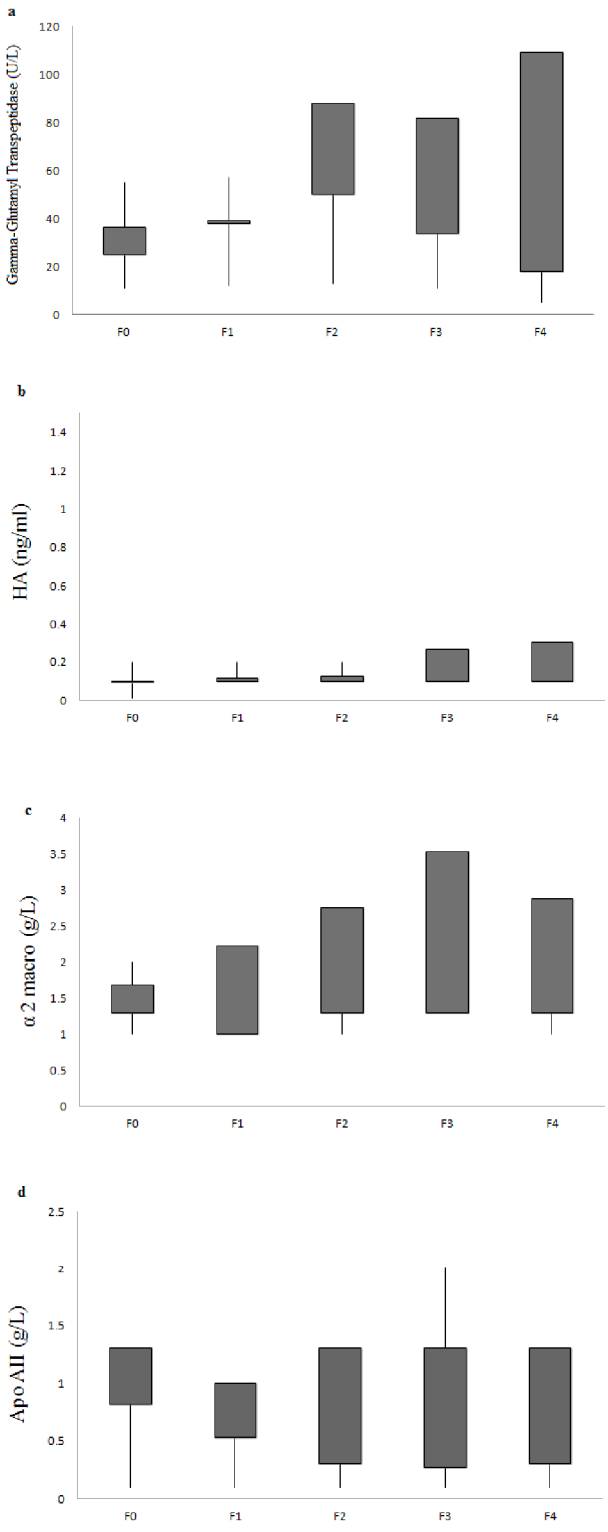

Fig. 3: Box-Whisker plots illustrating the distribution of values from (a) Gamma-Glutamyl Transpeptidase (GGT); (b) HA; (c) α 2-macroglobulin; and (d) ApoA2 with respect to fibrosis degrees. F0 represents no fibrosis, F1 represents portal fibrosis, F2 represents few septa, F3 represents many septa, and F4 represent cirrhosis.

Box plot diagram is used for visualization and illustrating the distribution selected values of the serum tests.

According to the calculations in the table above, nine of these parameters are statistically significant with a p-value <0.001. At 0.001 level of significance, ANOVA test showed that all of the nine fibrosis markers, differ significantly between the different stages of liver fibrosis as shown in Table above.

Taking an example of the variable HA, for HA there were significant differences between all pairwise combinations of fibrosis stages except between stage F0-F1 and F3-F4 (p <0.001) which means that later on F0 and F1 can be classified into one class and F3 and F4 can also be classified into one class (putting in mind that we are only considering one parameter).

VI. Classification: Decision Tree Approach

Classification is the problem of identifying to which category/set of categories a new case belongs to, based on the problem domain and the data included. In machine learning, classification is like identifying new sets by learning from old experiences [13].

In the categorization problem for classifying and predicting the fibrosis degrees (F0, F1, F2, F3, F4) instantaneously we used the Decision Tree classification technique.

As shown in Fig. 4 the inputs/variables are nine significant biomarkers which are liver function tests and other molecular tests: AST, ALT, ALB, T.BIL, D.BIL, GGT, HA, α 2-macroglobulin and ApoA2.

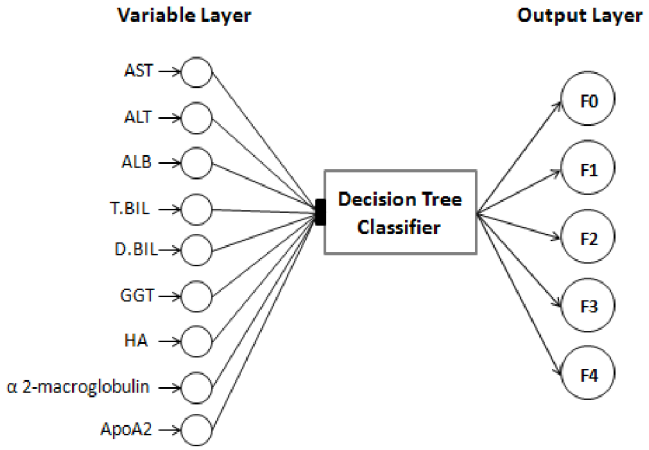

Fig. 4: Scheme for Single Stage Decision Tree classifier that classifying the five fibrosis degrees instantaneously.

Decision tree learning is one of machine learning approaches where it uses decision tree as a predictive model. Amongst other machine learning techniques decision tree is the easiest and simple to understand. It can easily handle both numerical and categorical data (compared to other techniques) [15].

A. Decision Tree classifier

After the significant variables are selected, we used decision tree classifier to set rules and constraints to the case study we have.

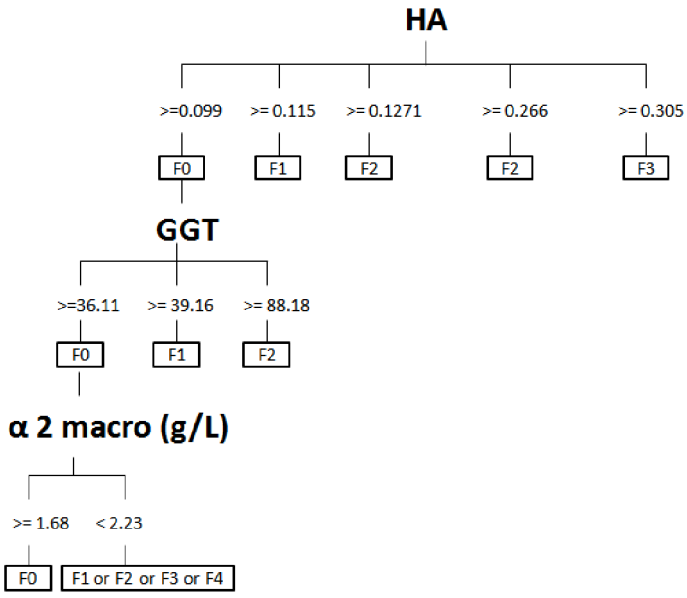

Fig. 5: Decision Tree for predicting F0 fibrosis "no" versus "significant, mild" fibrosis (F1, F2, F3, F4).

VII. Knowledge Acquisition and Prediction

The process of knowledge acquisition is considered as one of the main and critical process where data is transformed eventually into knowledge for prediction [6]. In this phase, we used decision tree as a machine learning techniques for classifying the data and setting the rules/constraints.

Not just linear analysis, but for finding correlations between parameters is a good indicator for accuracy. Equation 1 is used to measure the correlation between the proteins used in the data set, where association between variables are considered.

$$r = \frac{1}{n-1}\sum\left(\frac{x-x'}{S_x}\right)\left(\frac{y-y'}{S_y}\right) \quad (1)$$

After using the most significant biomarkers, mentioned in the feature selection section as model inputs to predict significant fibrosis (F1, F2, F3, or F4) versus "no" fibrosis (F0) and using the correlation equation with the decision tree classified, it has been proved that the classification accuracy is 93.7%.

As for instance, if there is a new patient with HCV virus and with the significant parameters calculations as within the ranges we already defined the stages of fibrosis: with a 93.7% the patient will be classified correctly.

Rule #1: IF Hyaluronic Acid (HA)>=0.099
THEN Fibrosis Class '0'

Rule #2: IF Gamma-Glutamyl Transpeptidase (GGT)>=36.11
AND Hyaluronic Acid (HA)>=0.099
THEN Fibrosis Class '0'

Rule #3: IF α2-macroglobulin (α2-MC)>=1.68
AND Gamma-Glutamyl Transpeptidase (GGT)>=36.11
AND Hyaluronic Acid (HA)>=0.099
THEN Fibrosis Class '0'

Rule #4: IF Hyaluronic Acid (HA)>=0.099
AND Gamma-Glutamyl Transpeptidase (GGT)>=36.11
AND α2-macroglobulin (α2-MC)<2.23
THEN Fibrosis Class '1 or 2 or 3 or 4'

Fig. 6: Rules/constraints from decision tree learning approach.

After using the decision tree classifier, a set of rules are concluded as shown in Fig. 5 and 6. Based on the selected variables the degree of fibrosis is predicted. For instance extracting the F0 degree "no fibrosis", variable 1 HA values range between 0-0.099 (ng/ml) AND variable 2 GGT values range between 36.11-255 (U/L) and so on for the rest of the variables.

Note that according to the parameters provided, various rules can be extracted and set based on the problem domain.

VIII. Medical Interpretation

The results obtained from this study demonstrated that a combination of nine biochemical markers (aspartateaminotransferase (AST) enzyme, alanineaminotransferase (ALT) enzyme, serum albumin (ALB), total bilirubin (T.BIL), direct bilirubin (D.BIL), Gamma-Glutamyl Transpeptidase (GGT), Hyaluronic Acid (HA), α 2-macroglobulin (α 2-MC) and apolipoprotine A2 (ApoA2)) analyzed with decision tree classifier have high positive predictive values for the diagnosis of different stages of fibrosis.

One of the main knowledge acquired was that HA level increases with increased fibrosis level and is superior in predicting different grades of fibrosis.

IX. Conclusion, discussion and future work

The importance of information system and strategic tools can never be denied. In this research, the statistical machine learning model for medical domain proved to have a positive effect on the research results and is very promising to help physicians with an analytical tool to improve the diagnostic processes with minimal need for medical procedures. The model can effectively and efficiently be further used for prediction and visualization of data to overcome the problem of on time decision for decision makers.

As well, by using the feature selection strategies, we could decrease the number of variables to be used by selecting the most significant variables which are: HA, GGT and α 2-MC. As the selected biomarkers showed statistical significant differences between fibrosis degrees by using FibroScan machine with (p-value $<$0.001).

In the future, other machine learning techniques can be used and compared to each others to easily predict the stage of the fibrosis. Furthermore, other biomarkers can be considered and calculated for significance as well as increasing the sample population.

References

[1] D.A. Saleh, F. Shebl, M. Abdel-Hamid, et al., "Incidence and risk factors for hepatitis C infection in a cohort of women in rural Egypt". Trans. R. Soc. Trop. Med. Hyg., vol. 102, pp. 921928, 2008.

[2] P. Bonnard, A. Elsharkawy, K. Zalata, E. Delarocque Astagneau, L. Biard, L. Le Fouler, A. B. Hassan, M. Abdel Hamid, M. El-Daly, M. E. Gamal, M. El Kassas, P. Bedossa, F. Carrat, A. Fontanet, and G. Esmat, "Comparison of liver biopsy and non-invasive techniques for liver fibrosis assessment in patients infected with HCV genotype 4 in Egypt". Journal of Viral Hepatitis, vol. 22, no. 3, pp. 245-253, 2015.

[3] F. El-Zanaty, Ann Way, Egypt Demographic and Health Survey 2008, Ministry of Health, El-Zanaty and Associates, and Macro International, Cairo, Egypt, March 2009.

[4] B. Cremilleux, N. Durand, Search for factors estimating the stage of liver fibrosis based on the discovery of meaningful clusters, in: PKDD 2002 Discovery Challenge on Hepatitis Data, Helsinki, Finland, 2002.

[5] G. Mahajani, Y. Aslandogan, Evidence Combination in Medical Data Mining, Department of Computer Science and Engineering University of Texas at Arlington, U.S.A. Technical Report, 2003.

[6] Ahmed M. Hashem, M. Emad M. Rasmy, Khaled M. Wahba and Olfat G. Shaker, "Single stage and multistage classification models for the prediction of liver fibrosis degree in patients with chronic hepatitis C infection". Comput Methods Programs Biomed., vol. 105, no. 3, pp. 194-209, 2012

[7] Heba Ayeldeen, Osman Hegazy and Aboul Ella Hassanien, "Case selection strategy based on K-means clustering". The 2nd International Conference on Information Systems Design and Intelligent Applications, Springer, India, 2015.

[8] Abdalla Zidan, Neveen I. Ghali, Aboul Ella Hassanien, Hesham Hefny, Jude Hemanth, "Level Setbased CT Liver Computer Aided Diagnosis System", Journal of Intelligent and Robotic Systems,Special Issue on Practical Perspective of Digital Imaging for Computational Applications, Volume 9, issue 1, 2013.

[9] Heba Ayeldeen, Olfat Shaker, Osman Hegazy and Aboul Ella Hassanien,"Distance similarity as a CBR technique for early detection of breast cancer: An Egyptian case study" in Information Systems Design and Intelligent Applications - Volume 340, pp 449-456, 2015.

[10] Ghany, K.K.A.; Hefny, H.A.; Hassanien, A.E.; Ghali, N.I., "A Hybrid Approach for Biometric Template Security," Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, pp. 941-942,2012.

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. MIT Press, 2012.

[12] Vladimir N. Vapnik. Statistical Learning Theory. Wiley- Interscience, 1998.

[13] Arthur, Samuel, "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal vol.3, no. 3, pp. 210229, 1959

[14] Wu, Q., and Zhou, D., Analysis of support vector machine classification, International Journal Computer Analysis Application, Vol. 8, pp. 99-119, 2006.

[15] Quinlan, J. R., "Induction of Decision Trees". Journal Machine Learning archive, vol. 1, no. 1, pp. 81-106, 1986,