Feature selection-based prediction of advanced fibrosis in HCV infected patients

ABSTRACT

Liver cirrhosis (LC), the end stage of chronic hepatitis C leads to liver cell failure and portal hypertension, which can favor the onset of hepatocellular carcinoma. Liver biopsy is considered as mandatory for the management of patients infected with the hepatitis C virus (HCV), particularly for staging of fibrosis degree. Nevertheless, the trend is to substitute the liver biopsy with non-invasive procedure, owing to its invasive nature and drawbacks of sampling error. The objective of this study is to evaluate different machine learning techniques and feature selection methods in prediction of advanced fibrosis by combining the serum biomarkers along with clinical information and histopathological findings to develop the classification models. The best developed classification model was able to predict the advanced fibrosis with accuracy of 52.73% for supervised Naïve Bayes model and 53.07% for unsupervised K-medoids model. This would pave the way to utilize classification models as a clinically non-invasive and reliable method to assess the degree of liver fibrosis.

INTRODUCTION

Hepatitis C is an infectious disease of the liver caused by the hepatitis C virus (HCV) [1]. Chronic hepatitis C (CHC) is recognized as a major healthcare problem worldwide and as a common infection in Egypt, especially genotype 4 and is a major global cause of chronic hepatitis, liver cirrhosis, and hepatocellular carcinoma [2]. Biologically, the end stage of chronic hepatitis C is a diffuse process characterized by fibrosis and the conversion of normal liver architecture into structurally abnormal nodules surrounded by annular fibrosis known as the Liver cirrhosis (LC). This persistent cumulative pathological disease results in liver cell collapse and portal hypertension, which may lead to hepatocellular carcinoma and death [3].

Assessment of liver fibrosis in Chronic Hepatitis C is mandatory for tracking the disease prognosis, determining the appropriate timing for medication, intervention approaches and predicting patient response [3,4]. The tools to estimate liver fibrosis may be invasive, such as liver biopsy, or non-invasive, which may be divided into serological tests and imaging instruments. Liver biopsy was considered as a gold standard in staging liver fibrosis [5] . 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]. 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 [7]. Also, numerous studies have shown that data mining is a powerful tool in the medical sector with great diagnostic potential of diseases due to its ability to discover the hidden predictive patterns from medical databases [9,10,11].

Several non-invasive serum markers and imaging techniques have emerged as non-invasive tools for diagnosis of fibrosis [8,11,12,13,14]. Serum markers are less invasive than biopsy, with no risk of complications, eliminate sampling and observer variability, easy to perform, and can be performed repeatedly [15]. The most common serum biochemical markers for HCV include aspartate aminotransferase (AST) and alanine aminotransferase (ALT) [16]. Other non-invasive methods in detection of fibrosis are based on indexes derived from serum markers[17], such as FIB-4 score and the aspartate aminotransferase (AST)-to-platelet ratio index (APRI) [18], or based on imaging techniques, such as using Transient Elastography (TE), which used ultrasound and vibratory waves for estimating the extent of liver fibrosis [19].

Studies show that the integration of clinical decision support and computer-based patient records with different machine learning classification techniques and features selection methods are useful for the prediction of different diseases [11]. In this study, I propose to identify the most relevant noninvasive biomarkers; provided from individual laboratory tests and findings of histological examinations Egyptian patients with HCV and compare and evaluate the usefulness of different machine learning techniques in prediction of advanced fibrosis.

Dataset Description

The goal of the study that generated the data was to develop, evaluate and validate a prediction model that replaces the invasive techniques, and to be a measurement to liver fibrosis progression. As well as to develop and validate computerized clinical decision-support system (CDSS) to support identification of individuals at higher risk of accelerated liver fibrosis progression [12]. Liver histology is determined via METAVIR score as assessed by local pathologists (including 13 centers around Egypt). According to the METAVIR system, fibrosis was staged on a scale from F0 toF4. F0 and F1 were considered as mild fibrosis, and F2, F3 and F4 as significant, whereas F3-F4 considered as advanced fibrosis [20]. The data contains reported clinical information regarding, but not limited, to the following: fever, nausea, fatigue, jaundice, gastric pain and weight loss depicted by body mass index(BMI) which are acute Hep C symptoms while age and gender influence the risk of progression to chronic infection[21]. Then it has the histological findings such as grade of fibrosis and the activity based on the gold standard, and laboratory tests such as alanine aminotransferase (ALT), as part ate aminotransferase (AST), white blood cells (WBC) count, red blood cells (RBC) count, Hemoglobin (Hb), platelet and quantity of HCV_RNA which help in determining the liver functioning. The RNA and ALT levels have been recorded during and post treatment (4 weeks, 12 weeks, end of treatment-post 12 weeks and efficacy) to justify the effectiveness of the treatment.

METHODS

The Dataset of 1385 Egypt HCV patients was accessed from UC Irvine Machine Learning Repository. Converting the .csv file into .xml using MS Excel was causing loss of data so, I used MATLAB to create a MATLAB data file and used the 'load *file.mat*' instruction to access it. There were no missing values. The output variable of stage was modified to 0 and 1 from a range of 1 to 4 to differentiate between moderate and advanced fibrosis [20] and all categorical variables were converted to binary form. The entire dataset was normalized using the min-max technique.

$$X_{new} = \frac{X_{old} - min(X_{old})}{\max(X_{old}) - \min(X_{old})}$$

Due to the presence of mixed data types, data visualization and statistical analysis was mostly performed individually for both data types.

Binary Data

Data visualization was done using heatmaps. It can be observed the cases are almost equally distributed in the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative

(FN) categories. Association with the output variable was determined using Chi-square test statistic. For significance level of 0.05, only the variable ‘NauseaVomting’ was found to be significant. Although, the heatmap does not explicitly differentiate the groups, the ttest does. All these patients were undergoing treatment for HCV. Their response to the treatment can be seen by observing the RNA levels and ALT levels at different point of time (Figure 3.3). And because the ALT levels kept fluctuating - the side effect of feeling nauseous is so strongly correlated to the output variable [21]. Bias corrected Cramer’s V was used to justify the strength of association of binary variables.

Continuous Data

The numeric data was visualized using histograms staged into the two prediction categories which tells us that irrespective of the stage the data is either right skewed or has no tapering values at the tails to show normality. Fixing the right-skew would contribute to a lot of distortion of the results and so it was not achieved. Spearman’s Rank Correlation was used to verify association with output variable and at significance level of 0.05, only the variable ‘BMI’ was found to be significant. This also supports the studies which say that BMI is supposed to inversely correlate with cirrhosis severity [22]. For feature selection, Kruskal Wallis test and Multinomial Regression has been used.

Feature Selection

Method	Features Selected
Supervised Learning	
TTest Based Categorical/Binary – Chi-Square Statistic Numeric – Multinomial regression	BMI, RNAEF, NauseaVomting
Step Based Both data types – Stepwise generalized linear model	BMI, NauseaVomting
Reduction Based Both data types – Logistic regression model	Age, Gender, BMI, Fever, NauseaVomting, Fatiguegeneralizedbodyache, Epigastricpain,WBC, RBC, Plat,AST1, ALT1, ALT4, ALT24,ALT48,ALTafter24w, RNABase, RNA4, RNA12, RNAEOT, RNAEF, BaselinehistologicalGrading
Unsupervised Learning	
PCA (PC1-2) Based	RNAEF, RNAEOT, RNA12, NauseaVomting, Plat, Epigastricpain, ALT1, ALT48, RBC, ALT24, WBC, RNA4
PCA (PC1-4) Based	RNAEF, RNAEOT, RNA12, NauseaVomting, Plat, Epigastricpain, ALT1, ALT48, RBC, ALT24, WBC, RNA4, HGB, Diarrhea, RNABase, BMI, Headache, ALT12, Jaundice, ALT4, Age

Table 2.1 Feature Selection methods used and features selected

ML techniques and metrics

6 machine learning algorithms were used, namely, logistic regression, tree bagger/random forest classification, multinomial regression, classification decision tree, lasso regularization and naïve bayes classification. The data was split into training and test data using the trainTestSplit function by 0.7 ratio. Metrics such as accuracy, precision and recall were calculated. Also, for decision trees the cvloss was calculated.

RESULTS

Data Exploration and Statistical Analysis

The acute Hep C symptoms are of categorical type and the rest are numeric type. Separate data visualization and statistical analysis was performed using MATLAB.

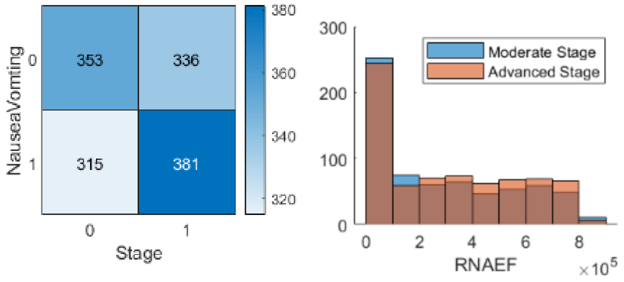


Figure 3.1,3.2 Showing Binary and Numeric Data Visualization

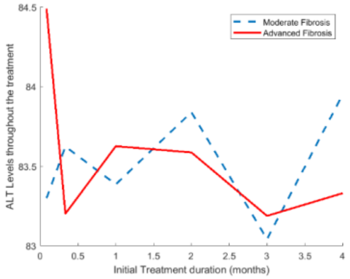
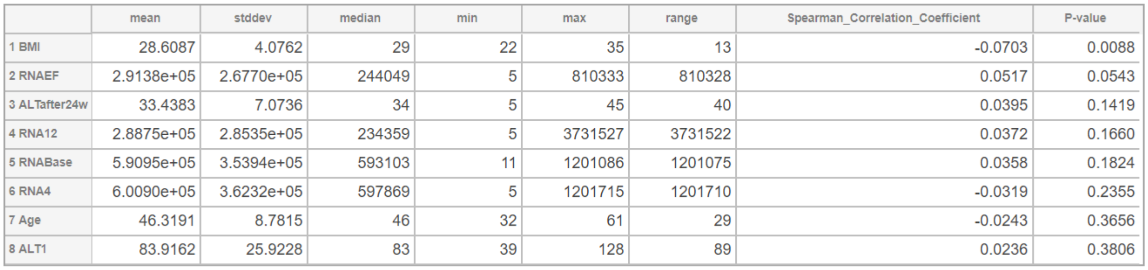
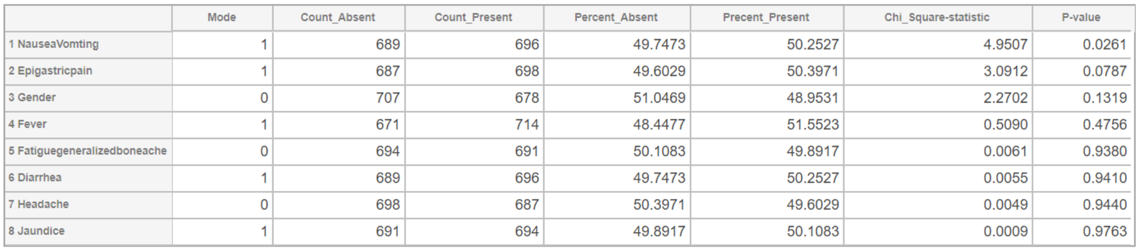


Figure 3.3 ALT levels timelapse



	mean	stddev	median	min	max	range	Spearman_Correlation_Coefficient	P-value
1 BMI	28.6087	4.0762	29	22	35	13	-0.0703	0.0088
2 RNAEF	2.9138e+05	2.6770e+05	244049	5	810333	810328	0.0517	0.0543
3 ALTafter24w	33.4383	7.0736	34	5	45	40	0.0395	0.1419
4 RNA12	2.8875e+05	2.8535e+05	234359	5	3731527	3731522	0.0372	0.1660
5 RNABase	5.9095e+05	3.5394e+05	593103	11	1201086	1201075	0.0358	0.1824
6 RNA4	6.0090e+05	3.6232e+05	597869	5	1201715	1201710	-0.0319	0.2355
7 Age	46.3191	8.7815	46	32	61	29	-0.0243	0.3656
8 ALT1	83.9162	25.9228	83	39	128	89	0.0236	0.3806

Table 3.1 Statistics of Top 8 Rank Correlated Numeric Features



	Mode	Count_Absent	Count_Present	Percent_Absent	Precent_Present	Chi_Square-statistic	P-value
1 NauseaVomting	1	689	696	49.7473	50.2527	4.9507	0.0261
2 Epigastricpain	1	687	698	49.6029	50.3971	3.0912	0.0787
3 Gender	0	707	678	51.0469	48.9531	2.2702	0.1319
4 Fever	1	671	714	48.4477	51.5523	0.5090	0.4756
5 Fatiguegeneralizedboneache	0	694	691	50.1083	49.8917	0.0061	0.9380
6 Diarrhea	1	689	696	49.7473	50.2527	0.0055	0.9410
7 Headache	0	698	687	50.3971	49.6029	0.0049	0.9440
8 Jaundice	1	691	694	49.8917	50.1083	0.0009	0.9763

Table 3.2 Statistics of Categorical Features

From Table 3.1, we can see that there is a large variation in the numeric/continuous values. Values for features like RBC, platelet and RNA levels are quite high (almost 10^5 times) than say age, BMI, AST, ALTs.

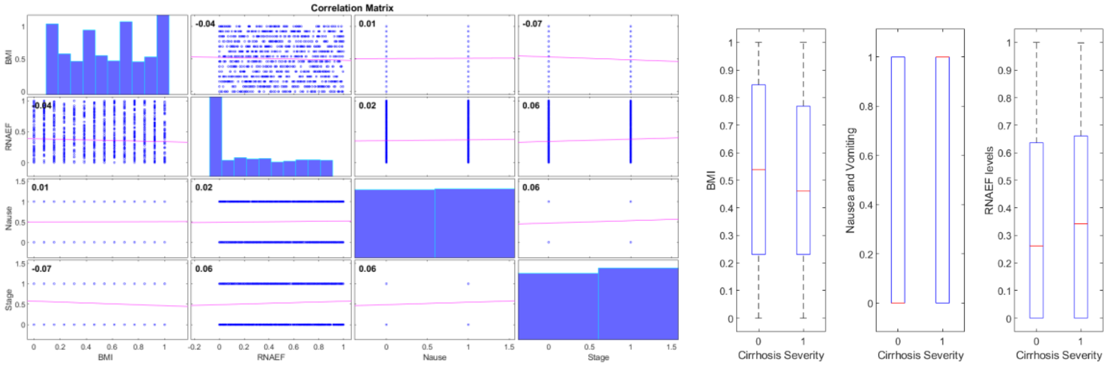


Figure 3.4 Correlation Matrix Plot and Boxplots of significant variables given by ttests

The differentiation in groups can be tested by performing two sample t-test. It also tells us the variable ‘RNAEF’ is also significant which contradicts the spearman correlation p-value. This is discussed later. The low correlation value can be explained by the slopes of the least-square regression lines of the correlation matrix plot above.

Unsupervised Learning Models

PCA was performed as a feature selection method while K-means and K-medoids were performed as binary classification models.

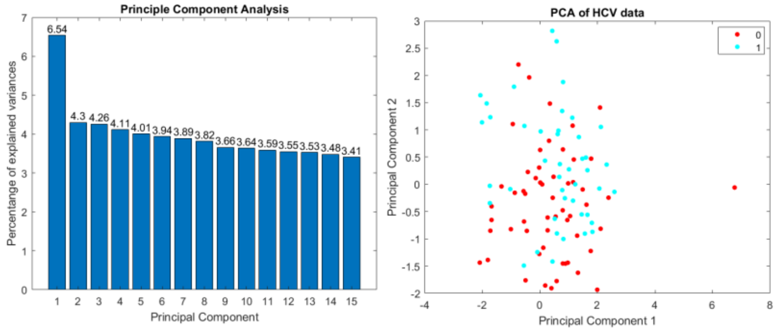
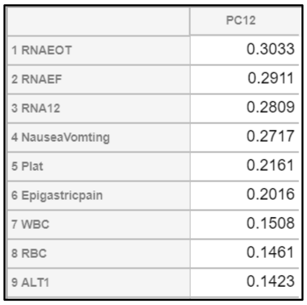
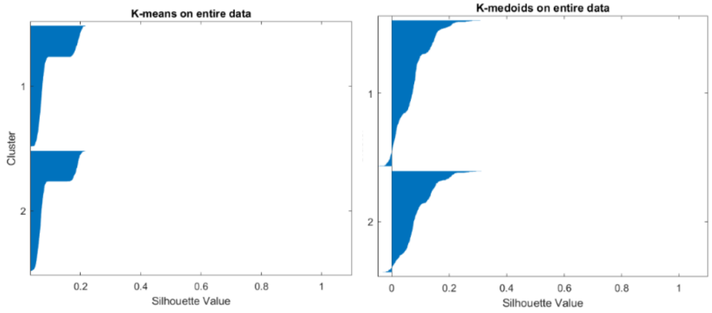


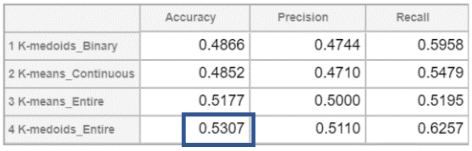
Figure 3.5 %variance explained by PCs and PC1 vs PC2



	PC12
1 RNAEOT	0.3033
2 RNAEF	0.2911
3 RNA12	0.2809
4 NauseaVomting	0.2717
5 Plat	0.2161
6 Epigastricpain	0.2016
7 WBC	0.1508
8 RBC	0.1461
9 ALT1	0.1423

Table 3.3 Top 9 variables of PC1-2



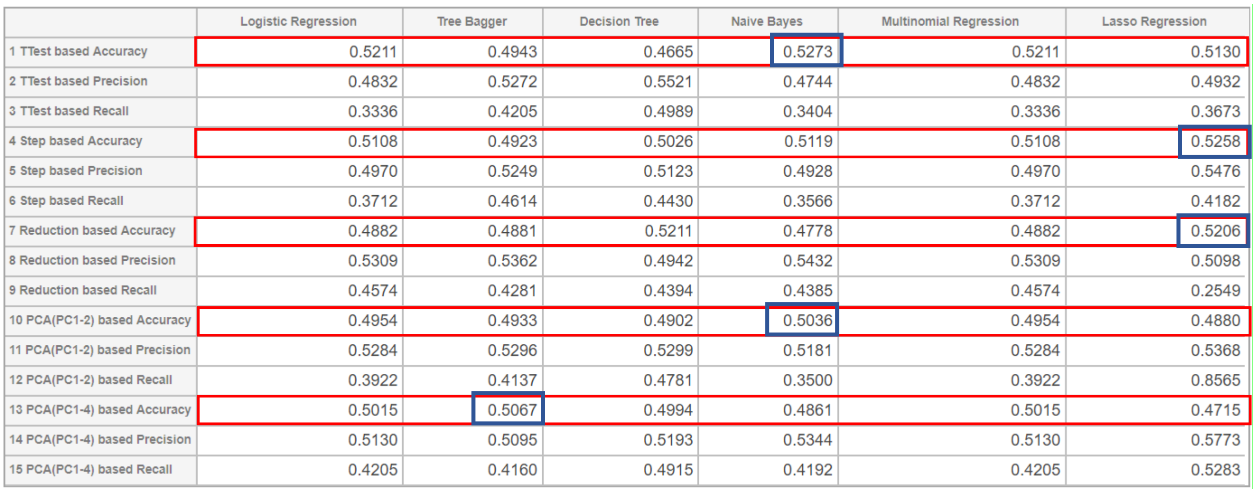


	Accuracy	Precision	Recall
1 K-medoids_Binary	0.4866	0.4744	0.5958
2 K-means_Continuous	0.4852	0.4710	0.5479
3 K-means_Entire	0.5177	0.5000	0.5195
4 K-medoids_Entire	0.5307	0.5110	0.6257

Table 3.4 Metrics of Unsupervised Binary classification - K-medoids performed on entire data achieves the maximum accuracy of 53.07%

Figure 3.6 K-means and K-medoids on entire data

Supervised Learning Models



	Logistic Regression	Tree Bagger	Decision Tree	Naive Bayes	Multinomial Regression	Lasso Regression
1 TTest based Accuracy	0.5211	0.4943	0.4665	0.5273	0.5211	0.5130
2 TTest based Precision	0.4832	0.5272	0.5521	0.4744	0.4832	0.4932
3 TTest based Recall	0.3336	0.4205	0.4989	0.3404	0.3336	0.3673
4 Step based Accuracy	0.5108	0.4923	0.5026	0.5119	0.5108	0.5258
5 Step based Precision	0.4970	0.5249	0.5123	0.4928	0.4970	0.5476
6 Step based Recall	0.3712	0.4614	0.4430	0.3566	0.3712	0.4182
7 Reduction based Accuracy	0.4882	0.4881	0.5211	0.4778	0.4882	0.5206
8 Reduction based Precision	0.5309	0.5362	0.4942	0.5432	0.5309	0.5098
9 Reduction based Recall	0.4574	0.4281	0.4394	0.4385	0.4574	0.2549
10 PCA(PC1-2) based Accuracy	0.4954	0.4933	0.4902	0.5036	0.4954	0.4880
11 PCA(PC1-2) based Precision	0.5284	0.5296	0.5299	0.5181	0.5284	0.5368
12 PCA(PC1-2) based Recall	0.3922	0.4137	0.4781	0.3500	0.3922	0.8565
13 PCA(PC1-4) based Accuracy	0.5015	0.5067	0.4994	0.4861	0.5015	0.4715
14 PCA(PC1-4) based Precision	0.5130	0.5095	0.5193	0.5344	0.5130	0.5773
15 PCA(PC1-4) based Recall	0.4205	0.4160	0.4915	0.4192	0.4205	0.5283

Table 3.5 Metrics of Supervised Learning Models for 5 fold cross validation

Naïve Bayes performed with TTest based feature selection method achieves the maximum accuracy of 52.73% and when its metrics and K-medoids on entire data's metrics are compared with random guess by ttest the recall value is considered as significant.

DISCUSSION AND CONCLUSION

From the statistical analysis of features(Table 3.1 and 3.2) based on p-value (significance level 0.05) the most ideal symptoms for advanced fibrosis seem to be suffering from nausea and vomiting and having low BMI. Except the PCA(PC1-2) feature selection method all other methods had selected these two variables. The two parameters (BMI and Nausea) found to be the most important features in prediction of the advanced fibrosis as they have statistically significant relationship (P-value < 0.05) with presence of advanced fibrosis as shown in the results. It was also justified that BMI negatively correlates with cirrhosis severity. The ALT levels variation could justify why nausea is a significant variable. Another benefit of Naïve Bayes model was found that it takes least time for execution, definitely because it is said to be the fastest model.

In this study, we made a comparison between different machine learning approaches on prediction of advanced liver fibrosis in Chronic Hepatitis C patients. Logistic Regression, Random Forest/Tree Bagger classification, Multinomial Regression, Classification Decision Tree, Lasso Regularization and Naïve Bayes Classification models were developed. We concluded that we could predict advanced fibrosis stage for chronic HCV patients using different machine learning approaches with maximum accuracy of 52.73% using supervised TTest based Naïve Bayes model and 53.07% using unsupervised K-medoids binary classification.

It is likely that we can improve the prediction of the advanced degree of fibrosis by taking the nature of temporal attributes into account, using APRI, AARI and FIB-4 scores calculated based on clinical information available. Also, it will be helpful to know the units of every feature, here they were assumed to be standard. Further approaches would be to explore artificial neural networks and other classification techniques.

REFERENCES

1. L. Gravitz, "A smouldering public-health crisis", *Nature*, vol. 474, no. 7350, pp. S2-S4, 2011.
2. El-Zanaty F, Ann Way. “Egypt Demographic and Health Survey 2008”. Cairo, Egypt:Ministry of Health, El-Zanaty and Associates,and Macro International, March 2009.
3. Anthony PP, Ishak KG, Nayak NC, Poulsen HE, Scheuer PJ, Sobin LH. The morphology of cirrhosis. Recommendations on definition, nomenclature, and classification by a working group sponsored by the World Health Organization. J Clin Pathol. 1978;31:395–414.
4. D. Crisan et al., "Prospective non-invasive follow-up of liver fibrosis in patients with chronic hepatitis C", *J. Gastrointest Liver Dis.*, vol. 21, pp. 375-382, 2012
5. Gebo K, Herlong H, Torbenson M, et al. “Role of liver biopsy in the management of chronic hepatitis C: A systematic review”. Hepatology, vol. 36, pp. 161-172, 2002.
6. P. Bedossa, F. Carrat, "Liver biopsy: The best not the gold standard", J. Hepatology, vol. 50, pp. 1-3, 2009.
7. H. Rosen, "Clinical practice Chronic Hepatitis C Infection", New England J. Med., vol. 364, no. 25, pp. 2429-2438, 2011
8. Soresi M, Giannitrapani L, Cervello M, Licata A, Montalto G. Non invasive tools for the diagnosis of liver cirrhosis. World J Gastroenterol. 2014 Dec 28;20(48):18131-50. doi: 10.3748/wjg.v20.i48.18131. PMID: 25561782; PMCID: PMC4277952.
9. B Mayer, H Rangwala, R Gupta et al., "Feature Mining for Prediction of Degree of Liver Fibrosis", AMIA Annu Symp Proc, pp. 1048, 2005.
10. Keltch, Brian, Yuan Lin, and Coskun Bayrak. ”Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients.” Journal of medical systems 38.8 (2014): 60.

11. S. El-Sappagh, F. Ali, A. Ali, A. Hendawi, F. A. Badria and D. Y. Suh, "Clinical Decision Support System for Liver Fibrosis Prediction in Hepatitis Patients: A Case Comparison of Two Soft Computing Techniques," in IEEE Access, vol. 6, pp. 52911-52929, 2018.
12. M. Nasr, K. El-Bahnasy, M. Hamdy and S. M. Kamal, "A novel model based on non invasive methods for prediction of liver fibrosis," 2017 13th International Computer Engineering Conference (ICENCO), Cairo, 2017, pp. 276-281.
13. S. Hashem et al., "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 3, pp. 861-868, 1 May-June 2018.
14. A. M. Hashem, M. E. M. Rasmy, K. M. Wahba and O. G. Shaker, "Prediction of the degree of liver fibrosis using different pattern recognition techniques," 2010 5th Cairo International Biomedical Engineering Conference, Cairo, 2010, pp. 210-214.
15. J. Parkes, I. N. Guha, P. Roderick, W. Rosenberg, "Performance of serum marker panels for liver fibrosis in chronic hepatitis C", J. Hepatol, vol. 44, pp. 462-474, 2006.
16. Gupta, Ekta & Bajpai, Meenu & Choudhary, Aashish. (2014). Hepatitis C virus: Screening, diagnosis, and interpretation of laboratory assays. Asian journal of transfusion science. 8. 19-25. 10.4103/0973-6247.126683.
17. S. Hashem et al., "A Simple multi-linear regression model for predicting fibrosis scores in chronic Egyptian hepatitis C virus patients", Int. J. Bio-Technol. Res., vol. 4, no. 3, pp. 37-46, Jun. 2014.
18. I. Zubair and B. Wajid, "Comparison of APRI, FIB-4 and fibro test in prediction of fibrosis and cirrhosis in patients with hepatitis C," 2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, 2018, pp. 222-227.
19. A. Wahba, N. Mohammed, A. Seddik, M. El-Adawy, "Liver fibrosis recognition using multi-compression elastography technique", J. Biomed. Sci. Eng. JBiSE, vol. 6, no. 11, pp. 1034-1039, 2013.
20. 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.
21. AAFP. (2010 Jun 1;81(11):1351-1357) Thad Wilkins, md; Jennifer k. Malcolm, do; Dimple Raina, md; and Robert r. Schade, md - *Hepatitis C: Diagnosis and Treatment*
22. Chaudhry, Asma et al. “To determine correlation between biochemical parameters of nutritional status with disease severity in HCV related liver cirrhosis.” Pakistan journal of medical sciences vol. 34,1 (2018): 154-158. doi:10.12669/pjms.341.14011