INTERIM PROJECT REPORT
April 11, 2020

BACKGROUND

Background of the Medical Problem

Chronic hepatitis C (CHC) is recognized as a major healthcare problem worldwide and as a common infection in Egypt, especially genotype 4. Hepatitis C virus (HCV) is a virus that infects liver cells and causes liver inflammation. Liver fibrosis occurs when the healthy tissue of liver becomes scarred and therefore cannot work as well. As more and more damage is done to the liver, scarring can occur. Fibrosis is the first stage of liver scarring. Later, if more of the liver becomes scarred, it's known as liver cirrhosis. Rates of progression to cirrhosis can vary dramatically across individuals. The assessment of liver fibrosis in CHC is essential to monitor the prognosis of the disease, to establish the optimal timing for therapy and management strategies, and to predict the response to treatment.

Liver biopsy is considered as mandatory for the management of patients infected with the hepatitis C virus (HCV), particularly for staging of liver fibrosis degree. Some call it the gold standard. However, liver biopsy has potential risk due its limitations including its invasive nature, being costly, being susceptible to sampling error, and the histological assessment that may suffer from variability of results. Therefore, in recent years the noninvasive methods have significantly increased in use as an alternative in staging chronic liver diseases for avoiding the drawbacks of biopsy based on serum markers and imaging techniques.

The serum markers of liver fibrosis offer an attractive alternative to liver biopsy; they are less invasive than biopsy, with no risk of complications, eliminate sampling and observer variability, and can be performed repeatedly unlike imaging where repeated tests increase costs.

How can AI help solve this in clinic?

Clinical risk prediction models in chronic hepatitis C virus (CHC) can be challenging due to non-linear nature of disease progression. 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. Thus, AI can help in predicting the non-linear nature of CHC.

State the objective of the study, what was the biological rationale for measuring the features used in the study?

The objective of the study is to consider both serum biomarkers and clinical information to evaluate machine learning technique and develop classification model for the prediction of advanced fibrosis in order to aid physicians in their decision-making process. The most common serum biochemical markers for HCV include aspartate aminotransferase (AST), alanine aminotransferase (ALT) and AST-to-platelet ratio (APRI) and fibrosis score (FIB-4).

Describe the data set - i.e. describe the features and observations

Dimensions of data: 1385 x 29

Feature Description:

- Features like fever, nausea, fatigue, jaundice, gastric pain and weight loss (depicted by BMI) are acute Hep C symptoms [5]
- The risk of progression to chronic infection is influenced by age and gender [4].
- The liver functioning can be determined by features such as WBC, RBC, Platelet, HB, AST, ALTs from the CBC [3].
- Histology results are the gold standards for classification.
- RNA levels during and post treatment (4 weeks, 12 weeks, end of treatment-post 12 weeks and efficacy) justify the effectiveness of the treatment.

Predictor variables are numeric as well as categorical –

- Features such as fever, nausea, fatigue, jaundice, gastric pain, diarrhea, gender and headache are categorical type.
- 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.

What is the objective of your machine learning (ML) algorithm?

The objective is to identify the best model and the selected features which play a key role in the prediction of HCV disease by using Hepatitis C Virus (HCV) from Egyptian patient's dataset from UCI in the content of multi and binary class labels.
Algorithms to be used for modeling are Multinomial logistic regression, Linear regression, Decision Tree, Random Forest, Artificial Neural Networks.

DATA EXPLORATION

Summary statistics - mean, median, range, missing values

There are no missing values.
Continuous Variables - For outlier detection one must consider only continuous variables. 4 outliers were found in the dataset. This was verified used boxplots and histograms. Statistics for these are stated below.

	Minimum	Maximum	Range	Mean	Median	Std.Dev.	Correlation
1 Age	32	61	29	46.3049	46	8.7717	-0.0176
2 BMI	22	35	13	28.5945	29	4.0731	-0.0584
3 WBC	2991	12101	9110	7.5388e+03	7514	2.6685e+03	0.0178
4 RBC	3816422	5018451	1202029	4.4223e+06	4438465	3.4657e+05	0.0106
5 HGB	10	15	5	12.5902	13	1.7135	0.0039
6 Plat	93013	226464	133451	1.5835e+05	157916	3.8817e+04	-0.0174
7 AST1	39	128	89	82.7581	83	25.9888	-0.0240
8 ALT1	39	128	89	83.9240	83	25.9375	0.0389
9 ALT4	39	128	89	83.4359	83	26.5488	-0.0163
10 ALT12	39	128	89	83.4967	84	26.0676	-0.0013
11 ALT24	39	128	89	83.6524	83	26.2091	-0.0052
12 ALT36	39	128	89	83.2795	84	26.1830	-0.0015
13 ALT48	39	128	89	83.7958	84	26.0037	-0.0088
14 ALTafter24w	22	45	23	33.5025	34	6.9577	0.0404
15 RNABase	11	1201086	1201075	5.9118e+05	593103	3.5425e+05	0.0309
16 RNA4	5	1201715	1201710	6.0041e+05	595314	3.6263e+05	-0.0339
17 RNA12	5	810028	810023	2.8668e+05	234359	2.7009e+05	0.0512
18 RNAEOT	5	808450	808445	2.8772e+05	251376	2.6444e+05	-0.0183
19 RNAEF	5	810333	810328	2.9164e+05	244418	2.6776e+05	0.0301
20 BaselinehistologicalGrading	3	16	13	9.7589	10	4.0232	-0.0465
21 Baselinehistologicalstaging	1	4	3	2.5358	3	1.1208	1.0000

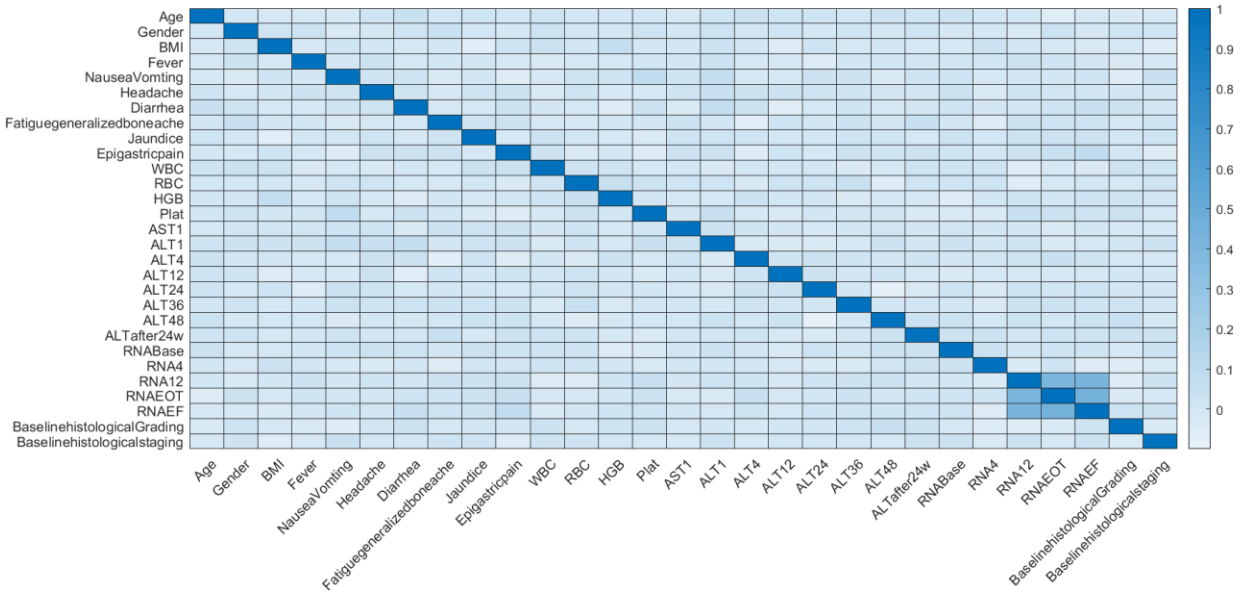
Categorical Variables – Number of cases

	Stage 1	Stage 2	Stage 3	Stage 4	Total
1 Male	172	181	162	190	705
2 Female	164	149	192	171	676
3 Fever Absent	157	159	166	186	668
4 Fever Present	179	171	188	175	713
5 NausaeVomiting Absent	180	171	165	170	686
6 NausaeVomiting Present	156	159	189	191	695
7 Headache Absent	167	168	180	181	696
8 Headache Present	169	162	174	180	685
9 Diarrhea Absent	160	171	178	178	687
10 Diarrhea Present	176	159	176	183	694
11 Fatigue Absent	175	158	183	176	692
12 Fatigue Present	161	172	171	185	689
13 Jaundice Absent	179	154	182	176	691
14 Jaundice Present	157	176	172	185	690
15 GastricPain Absent	147	168	187	184	686
16 Gastric Pain Present	189	162	167	177	695

Statistical analysis

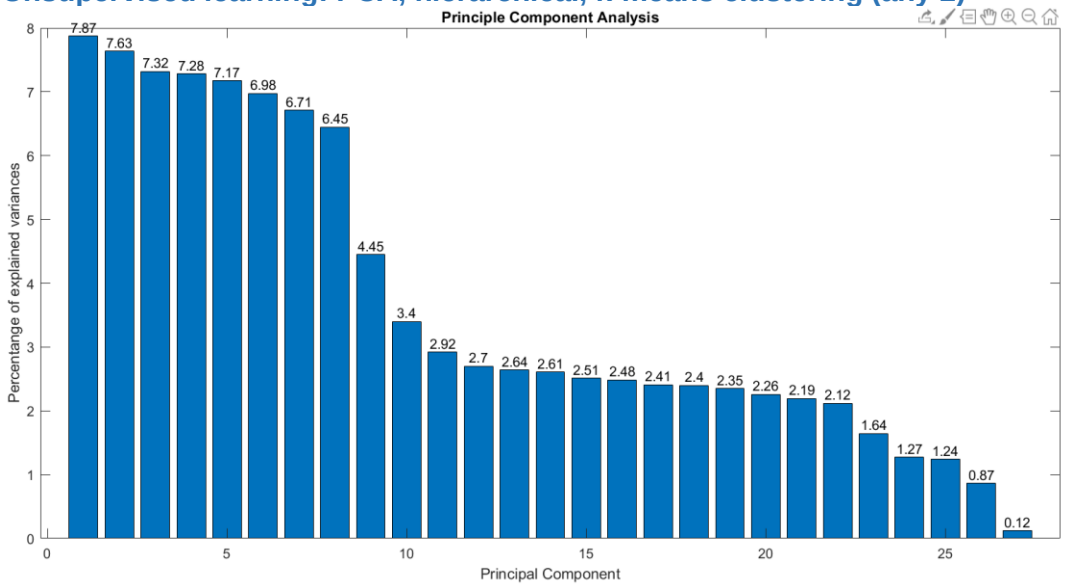
	Pearson_Pval	Pearson_Qval	Pearson_FDR	Anova_Pval	Anova_Qval	Anova_FDR
1 Age	0.4921	13.2879	0.6142	0.3448	9.3088	1.0000
2 Gender	0.7095	19.1560	0.5755	0.1010	2.7279	0.6476
3 BMI	0.0359	0.9705	0.5831	0.0094	0.2531	0.2404
4 Fever	0.2687	7.2544	0.4843	0.5496	14.8393	1.0000
5 NauseaVomting	0.0410	1.1079	0.3329	0.1765	4.7642	0.9048
6 Headache	0.9216	24.8833	0.5751	0.9870	26.6486	0.9372
7 Diarrhea	0.7782	21.0114	0.5739	0.7432	20.0675	0.9528
8 Fatiguegeneralizedboneache	0.6045	16.3216	0.5162	0.6153	16.6119	0.9279
9 Jaundice	0.4642	12.5329	0.6276	0.3328	8.9850	1.0000
10 Epigastricpain	0.0526	1.4189	0.2842	0.0871	2.3522	0.7445
11 WBC	0.5241	14.1508	0.5002	0.7665	20.6963	0.8933
12 RBC	0.7756	20.9410	0.5992	0.6444	17.4000	0.9179
13 HGB	0.8966	24.2082	0.5818	0.7643	20.6360	0.9331
14 Plat	0.5151	13.9078	0.5571	0.5917	15.9761	1.0000
15 AST1	0.3554	9.5949	0.5241	0.4598	12.4147	1.0000
16 ALT1	0.1584	4.2766	0.6424	0.4407	11.8980	1.0000
17 ALT4	0.5204	14.0505	0.5277	0.6772	18.2849	0.9138
18 ALT12	0.9922	26.7893	0.5962	0.9533	25.7386	0.9400
19 ALT24	0.8118	21.9176	0.5487	0.8996	24.2884	0.9225
20 ALT36	0.7904	21.3402	0.5575	0.8969	24.2165	0.9581
21 ALT48	0.5931	16.0143	0.5346	0.5961	16.0936	0.9551
22 ALTafter24w	0.2201	5.9424	0.5101	0.5283	14.2630	1.0000
23 RNABase	0.2657	7.1740	0.5388	0.3713	10.0251	1.0000
24 RNA4	0.2100	5.6692	0.5677	0.4954	13.3760	1.0000
25 RNA12	0.1979	5.3439	0.6422	0.5648	15.2489	1.0000
26 RNAEOT	0.4924	13.2957	0.5706	0.8797	23.7513	0.9806
27 RNAEF	0.2918	7.8795	0.4735	0.0840	2.2682	1.0000

Ancova might give a better understanding.



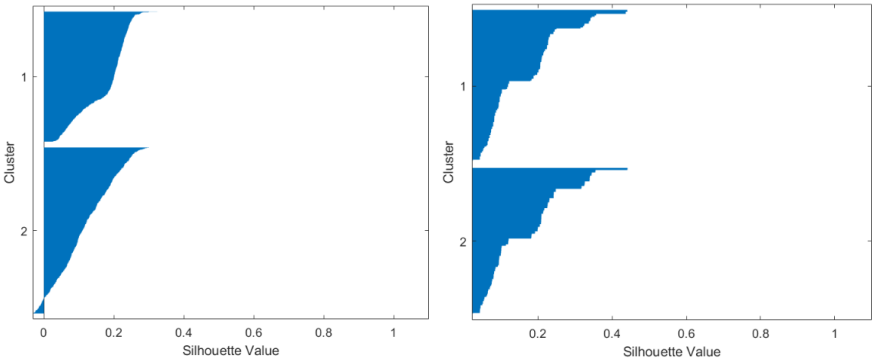
Supported by MRMR (Minimum Redundancy Maximum Relevance) algorithm developed for ranking features of a mixed dataset.

Unsupervised learning: PCA, hierarchical, k-means clustering (any 2)



K-means for continuous variables

K-medoids for categorical



INTERPRETATION

Interpret results from data exploration, propose hypothesis for the supervised ML model

Subset of features needs to be selected for proper analysis and to improve the accuracy. RNA levels appear to be significant features. Although on observing unsupervised classification, it is possible to classify the given dataset into moderate fibrosis and advanced fibrosis. Advanced fibrosis can be determined using serum biomarkers and clinical information is the hypothesis for the supervised ML model.

ML

Apply 1 supervised ML algorithm (like Linear or logistic regression) and test using hold out/cross validation

Aim1 – Used multinomial logistic regression to predict stage of liver fibrosis using serum biomarkers and clinical information.

Improvised based on findings from unsupervised learning.

Aim2 – Used multinomial logistic regression, linear regression to predict advanced liver fibrosis.

Also used subset of features obtained from stepwiselm and observed that the precision, accuracy and recall increased as compared to using all features.[5]

REFERENCES

1. IEEE Xplore. (12 February 2018) Mei-Hua Tsai, Kuei-Hsiang Lin, Kuan-Tsou Lin, Chi-Ming Hung, Hung-Shiang Cheng, Yu-Chang Tyan, Hui-Wen Huang, Bintou Sanno-Duanda, Ming-Hui Yang, Shyng-Shiou Yuan and Pei-Yu Chu - *Predictors for Early Identification of Hepatitis C Virus Infection* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4564624/
2. HepatitisC.net - *Understanding Lab Test Results* https://hepatitisc.net/diagnosis/understanding-lab-test-results/

3. Asian Journal of Transfusion Science. (2014 Jan-Jun; 8(1): 19–25. doi: 10.4103/0973-6247.126683) Ekta Gupta, Meenu Bajpai, and Aashish Choudhary - *Hepatitis C virus: Screening, diagnosis, and interpretation of laboratory assays* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3943138/

4. 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* https://www.aafp.org/afp/2010/0601/p1351.html
5. https://www.mathworks.com/help/stats/feature-selection.html