Multiclass Classification of Hepatitis-C, Cirrhosis and Fibrosis Using Multiple Classifier Models

1st Maidul Islam
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
maidul.islam@g.bracu.ac.bd

2nd Farjana Alam
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
farjana.alam@g.bracu.ac.bd

3rd Md Sabbir Hossain
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
md.sabbir.hossain1@g.bracu.ac.bd

4th Farah Binta Haque
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
farah.binta.haque@g.bracu.ac.bd

5th Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Hepatitis C Virus (HCV) is a blood-borne illness that is infectious and first mostly asymptomatic. As a result, it is challenging to identify and manage individuals who are in the early stages of infection. As the disease advances to its latter stages, diagnosis and therapy become increasingly challenging. In this paper, a machine learning-based AI system for assisting medical practitioners in the early identification of hepatitis C is provided. The chance of people contracting HCV infection may be predicted using medical data such as age, sex, place of residency, and (ALT, AST, ALB, ALP) enzyme blood tests. It may be used as a teaching tool to teach nurses and medical students how to identify people with HCV infection. The efficacy of several classification algorithms for predicting HCV infection from the HCV data set is examined in this study. SVM, AdaBoost, Logistic Regression, KNN and Random Forest are some of these techniques. Results indicate that each approach has a special advantage in achieving the specified classification. Additionally, we also discussed briefly about the best classification algorithm by their classification reports such as accuracy, f1 score, precision, recall, confusion matrix etc.**

***Index Terms*—Hepatitis-C, SVM, KNN, Random Forest, Naïve Bayes, Logistic Regression, AdaBoost, Confusion Matrix, Accuracy.**

I. Introduction

The liver can get infected by the hepatitis C virus. Acute and chronic hepatitis, which can range in severity from a moderate condition to a serious, life-long condition involving liver cirrhosis and cancer, can both be brought on by the virus. It is a bloodborne virus, and most infections result through contact with blood via risky injection techniques, risky medical procedures, unscreened blood transfusions, injectable drug usage, and sexual behaviors that expose one to blood.

Acute and chronic infection are both brought on by the hepatitis C virus (HCV). Most acute HCV infections are asymptomatic and do not progress to a serious condition that threatens life. Within 6 months of infection, 30% (15–45%) of infected individuals naturally remove the virus without receiving any therapy. The remaining 70% (55–85%) of people will become infected with HCV chronically. Cirrhosis can develop in people with persistent HCV infection at a rate of 15% to 30% within 20 years.

We explored various classification models and found that the Logistic Regression model performed the best. In terms of the multi class classifications, we can see that the Logistic Regression model outperforms many of the other learners and algorithms that we chose for comparison.

II. Methodology

In this study, we used several classification algorithms like Support Vector Machine (SVM), Random Forest classifier, KNN (K-Nearest Neighbor), AdaBoost, Naive Bayes classifier, Logistic Regression to classify different categories such as Blood Donor, Suspect Blood Donor, Hepatitis, Fibrosis and Cirrhosis.

A. Random Forest:

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues.

On various samples, it constructs decision trees and uses their average for classification and majority vote for regression.

B. KNN Algorithm:

The K-nearest neighbors (KNN) method predicts the values of new data points based on "feature similarity," which further indicates that the new data point will be given a value depending on how closely it resembles the points in the training set.

C. SVM Algorithm:

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression issues. The SVM algorithm's objective is to establish the optimal line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future.

D. AdaBoost Algorithm:

The abbreviation "AdaBoost" stands for "Adaptive Boosting," which is a highly well-liked boosting approach that turns several "poor classifiers" into one "strong classifier."

E. Naive Bayes:

The Bayes Theorem is the foundation of the probabilistic machine learning method known as Naive Bayes, which is utilized for a variety of classification problems.

F. Logistic Regression:

In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. Rather than providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false.

III. Dataset Description

In the HCV dataset, there were 615 instances for every attributes. The dataset we used for our analysis was taken from UCI Machine Learning Repository. The dataset contained 14 attributes named as X (Patient ID/No.), Category (0=Blood Donor, 0s=suspect Blood Donor, 1=Hepatitis, 2=Fibrosis, 3=Cirrhosis), Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT and PROT. All the attributes except 'X' were taken into consideration while applying the classification algorithms. Attributes named ALB, ALP, CHOL, PROT and ALT had missing values in some of their instances. The Category attribute had ordinal categorical values and the Sex attribute had nominal categorical values. All the other attributes had numeric values in their instances.

IV. Data Preprocessing

For data preprocessing we went through five crucial stages such as Dropping unnecessary columns, missing value imputation, Encoding the columns containing categorical values.

- **Dropping Unnecessary Column:**The column X had no correlation with the dependent variable y (Category column, in this case). So, we considered this column as unnecessary and removed the column.

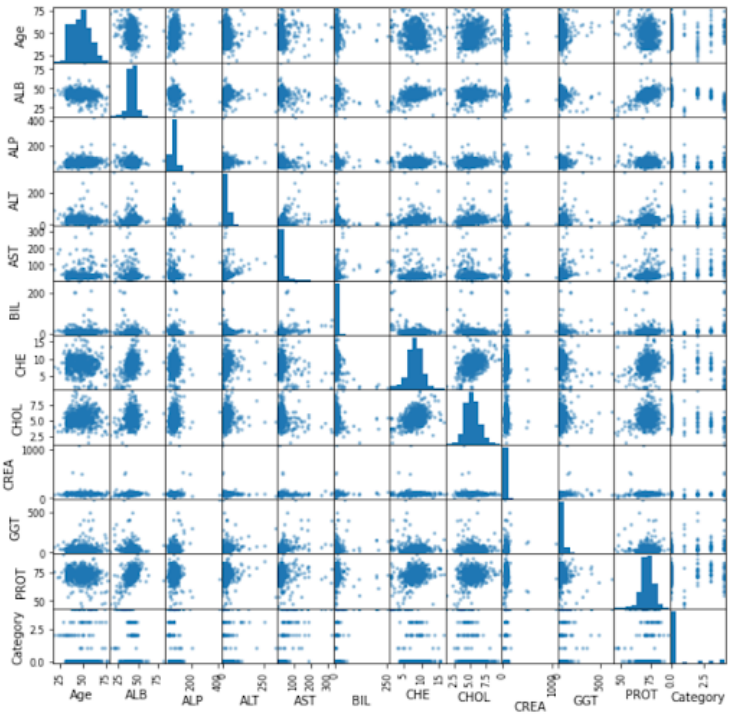

Fig. 1. Scatter Matrix of the dataset

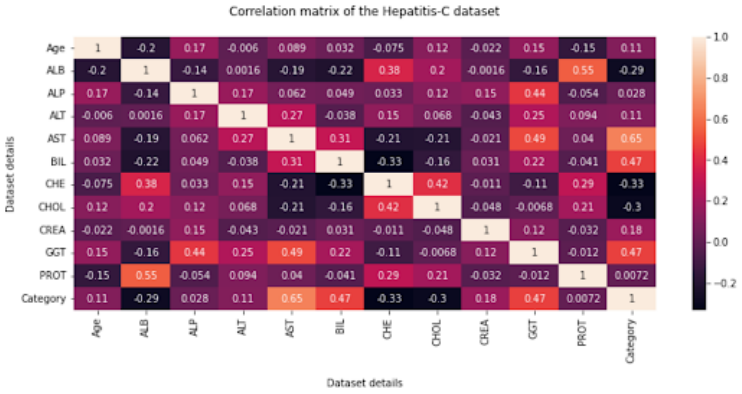

Fig. 2. Correlation matrix of the Hepatitis-C dataset.

- **Missing value imputation:**In deep learning and machine learning, null values are a major issue. Before feeding our data to machine learning algorithms, it is important to handle the missing values. The ALB, ALP, CHOL, PROT and ALT had missing values in some of their instances. So, we took the mean of every attribute and placed them instead of the missing values or Null values. Thus, our applied machine learning algorithms overcame the risk of getting unpleasant errors.
- **One hot encoding the Sex column:**One hot encoding better works with data containing nominal categorical values. Data can be converted using one hot encoding as a means of getting a better prediction and preparing the data for an algorithm. With one-hot, we create a new category

column for each categorical value and give it a binary value of 1 or 0. First of all, we removed the original Sex column and created two different columns using One hot encoding named Sex_m and Sex_f which represents the gender of the person containing binary values.

- **Label Encoding the Category column:**Label encoding is the process of transforming labels into a numeric form so that they may be read by machines. It is often used with labels containing ordinal values. The operation of such labels can then be better determined by machine learning techniques. It is an important preprocessing step for the structured dataset in machine learning. As the category column had ordinal values, we converted the labels like this, Blood Donor=0, suspect Blood Donor=1, Hepatitis=2, Fibrosis=3, Cirrhosis=4.

V. Experiment

6 different types of algorithms such as Support Vector Machine (SVM), Random Forest classifier, KNN (K-Nearest Neighbor), AdaBoost, Naive Bayes classifier and Logistic Regression were used in the dataset. All the algorithms worked very well while performing the classifications. To evaluate the performance of the experiment, the best way is to get the confusion matrix. Through confusion matrix it is possible to detect the type of errors made by the classifiers.

- **Random forest Classifier:**With the Random Forest classifier, among the 615 instances, 97% of the Blood donor class were correctly classified. All the instances of Hepatitis class were correctly classified. Fibrosis and Cirrhosis both showed an accuracy of 50% using the Random Forest algorithm.

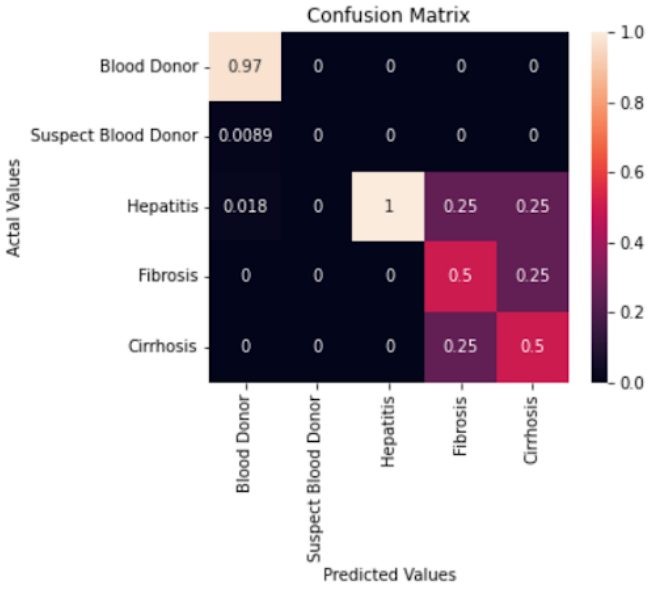

Fig. 3. Confusion matrix for Random forest

- **K-Nearest Neighbors (KNN):**With the K-Nearest Neighbors (KNN) classifier, among the 615 instances, 96% of the Blood donor class were correctly classified. On the other hand,75% of the Hepatitis class were correctly classified. Fibrosis and Cirrhosis showed an accuracy of 20% and 50% respectively by using the K-Nearest Neighbors (KNN) algorithm.

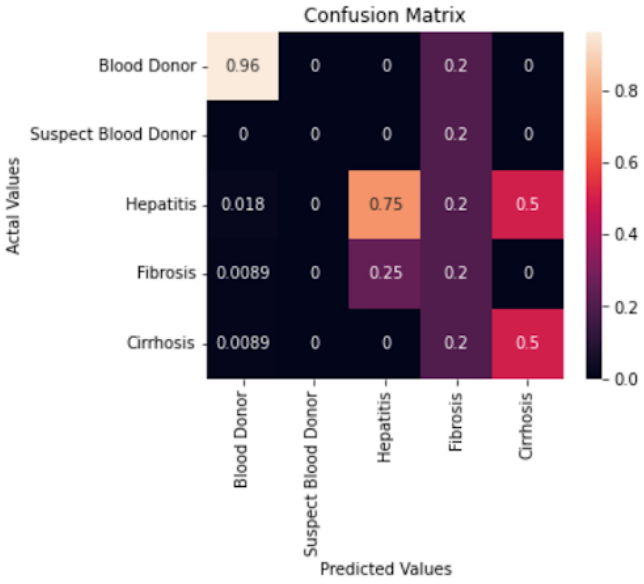

Fig. 4. Confusion matrix for K-Nearest Neighbors

- **Support Vector Machine (SVM):**With the Support Vector Machine (SVM) classifier, among the 615 instances, 96% of the Blood donor class were correctly classified. There were slight errors while classifying the Suspected Blood Donor and Fibrosis classes. On the other hand,75% of the Hepatitis class were correctly classified. Hepatitis and Cirrhosis both showed an accuracy of 33% using the Support Vector Machine (SVM) algorithm.
- **Naive Bayes Classifier:**With the Naive Bayes Classifier, among the 615 instances, 97% of the Blood donor class were correctly classified. On the other hand, Hepatitis, Cirrhosis and Fibrosis showed 20%, 22% and 67% accuracy with the Naive Bayes Classifier.
- **AdaBoost:**With the AdaBoost algorithm, among the 615 instances, 97% of the Blood donor class were correctly classified. On the other hand, Hepatitis and Fibrosis showed 33% and 25% accuracy respectively. Moreover, all of the instances of the Cirrhosis class were perfectly classified using the AdaBoost algorithm.
- **Logistic Regression:**In terms of Logistic Regression, among the 615 instances, 97% of the Blood donor class were correctly classified. On the other hand, Hepatitis showed 80% accuracy. Moreover, among all of the instances of the Cirrhosis class and Fibrosis class were classified with 50% accuracy using the Logistic Regression algorithm.

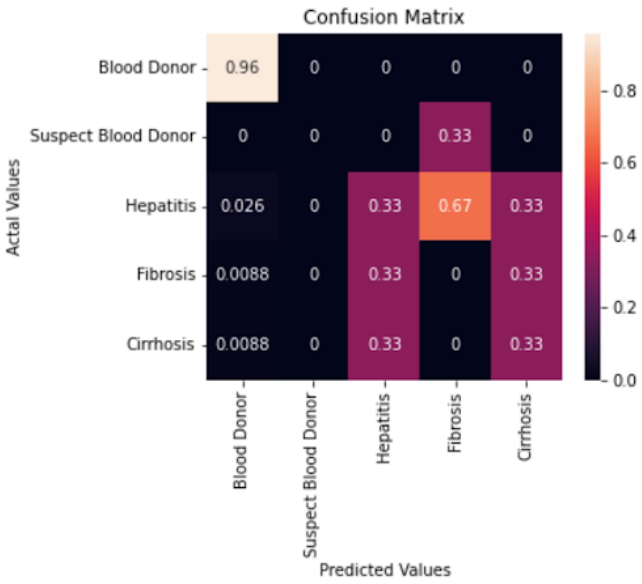

Fig. 5. Confusion matrix for Support Vector Machine

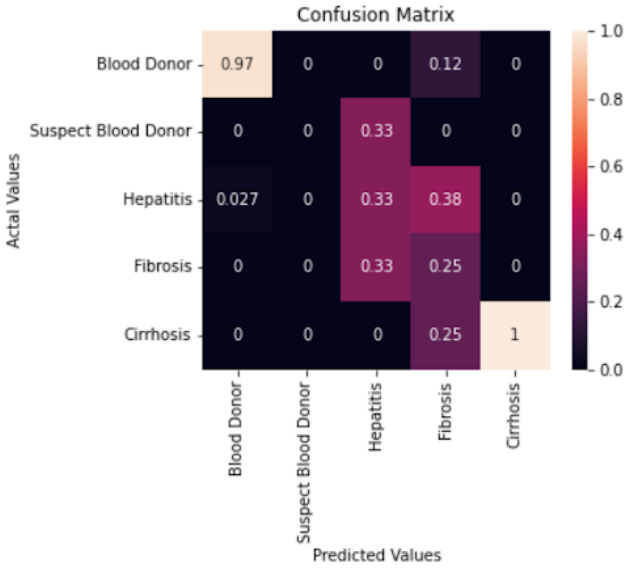

Fig. 7. Confusion matrix for AdaBoost

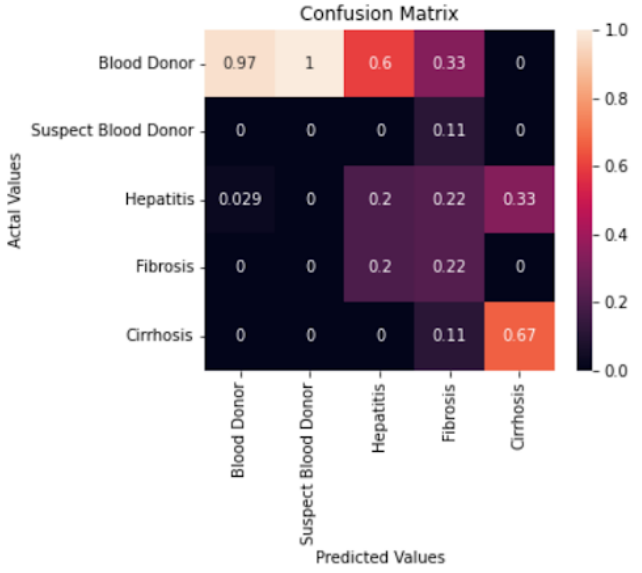

Fig. 6. Confusion matrix for Naive Bayes

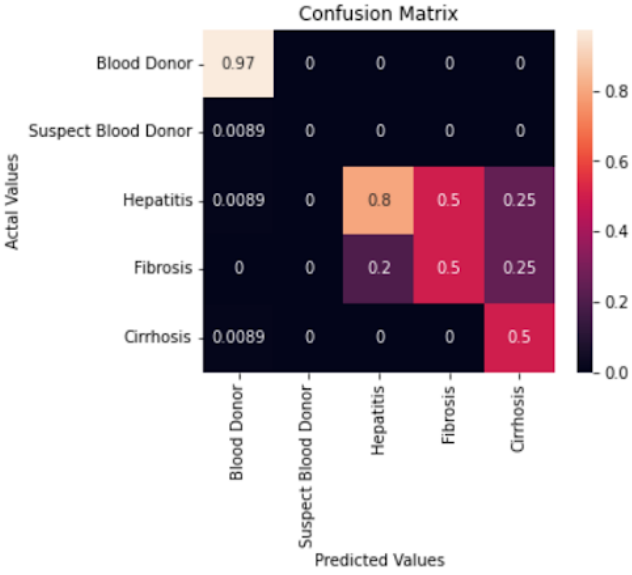

Fig. 8. Confusion matrix for Logistic Regression

RESULTS

- **Random Forest Classifier:**
- **KNN:**
- **SVM:**
- **Naïve Bayes:**
- **AdaBoost Classifier:**
- **Logistic Regression:**

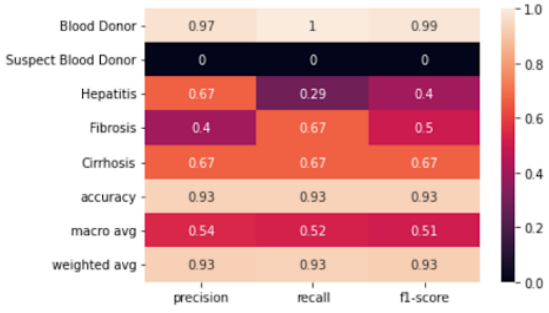

Fig. 9. Confusion matrix for AdaBoost

REFERENCES

[1]Berg, T., Sarrazin, C., Herrmann, E., Hinrichsen, H., Gerlach, T., Zachoval, R., ... & Zeuzem, S. (2003). Prediction of treatment outcome in patients with chronic hepatitis C: significance of baseline parameters

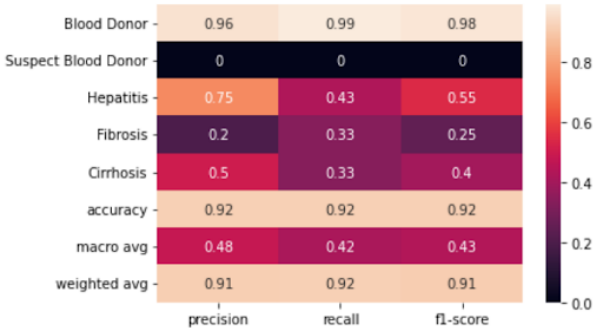

Fig. 10. Result of KNN

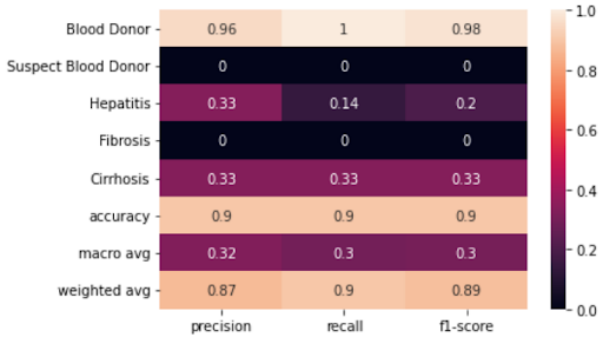

Fig. 11. Result of SVM

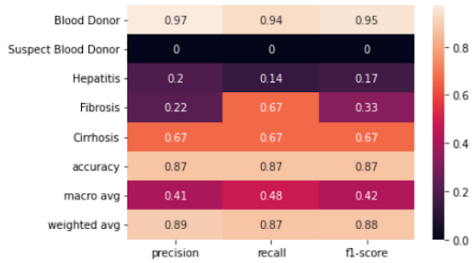

Fig. 12. Result of SVM

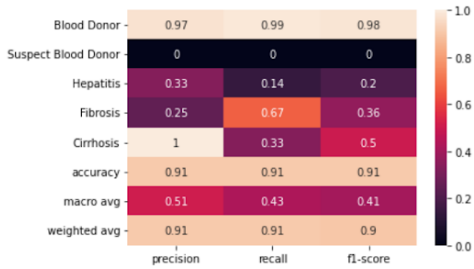

Fig. 13. Result of AdaBoost Classifier

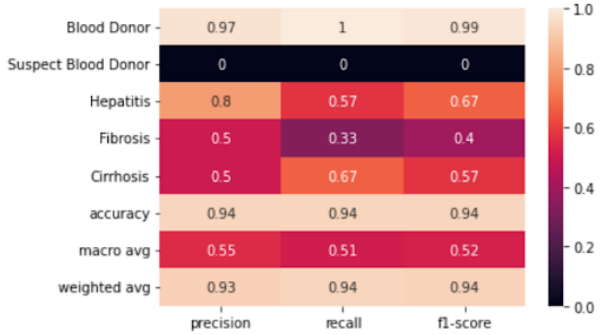

Fig. 14. Result of Logistic Regression

and viral dynamics during therapy. Hepatology, 37(3), 600-609.
[2]Zou, S., Tepper, M., & El Saadany, S. (2000). Prediction of hepatitis C burden in Canada. Canadian Journal of Gastroenterology, 14(7), 575-580.
[3]Barakat, N. H., Barakat, S. H., & Ahmed, N. (2019). Prediction and staging of hepatic fibrosis in children with hepatitis c virus: A machine learning approach. Healthcare Informatics Research, 25(3), 173-181.
[4]Ahammed, K., Satu, M. S., Khan, M. I., & Whaiduzzaman, M. (2020, June). Predicting infectious state of hepatitis c virus affected patient's applying machine learning methods. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 1371-1374). IEEE.
[5]ElHefnawi, M., Abdalla, M., Ahmed, S., Elakel, W., Esmat, G., Elraziky, M., ... & Hassan, M. (2012, August). Accurate prediction of response to Interferon-based therapy in Egyptian patients with Chronic Hepatitis C using machine-learning approaches. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 771-778). IEEE.
[6]Ioannou, G. N., Tang, W., Beste, L. A., Tincopa, M. A., Su, G. L., Van, T., ... & Waljee, A. K. (2020). Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. JAMA network open, 3(9), e2015626-e2015626. [7]Akella, A., & Akella, S. (2020). Applying machine learning to evaluate for fibrosis in chronic hepatitis c. medRxiv.