Predicting Acute Liver Failure using Clinical and Histopathological Data

A PROJECT REPORT
Submitted as part of BBD451 BTech Major Project

Submitted by:
Yukti Makhija
2019BB10067

Guided by:
Prof. Ishaan Gupta

DEPARTMENT OF BIOCHEMICAL ENGINEERING AND BIOTECHNOLOGY
INDIAN INSTITUTE OF TECHNOLOGY DELHI

September 2022

DECLARATION

I certify that

a) the work contained in this report is original and has been done by me under the guidance of my supervisor(s).

b) I have followed the guidelines provided by the Department in preparing the report.

c) I have conformed to the norms and guidelines given in the Honor Code of Conduct of the Institute.

d) whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources whenever necessary.

Signed by: Yukti Makhija (2019BB10067)

CERTIFICATE

It is certified that the work contained in this report titled “Predicting acute liver failure using clinical and histopathological data” is the original work done by Ms Yukti Makhija and has been carried out under my supervision.

Signed by: Prof. Ishaan Gupta, IITD
Date: 23/09/22

TABLE OF CONTENTS

DECLARATION 1

CERTIFICATE 2

TABLE OF CONTENTS 3

List of Tables 4

List of Figures 5

List of Abbreviations 6

Chapter 1: Introduction 7

Objectives 7

Chapter 2: Literature Review 8

Chapter 3: Methods 9

Data Description 9

Types of CLD 9

Preprocessing the Dataset 10

Parser to Extract Information from EMRs 10

Chapter 4: Results 12

Projection and Clustering 12

Logistic Regression 13

Decision Tree 13

Support Vector Machine 15

Chapter 5: Conclusion 17

Next Steps 17

References 18

List of Tables

Table	Title	Page
1	Dataset diagnosis distribution	9
2	Dataset after preprocessing, descriptions of parameters	10
3	Logistic Regression performance metrics	13
4	Decision Tree performance metrics	14
5	SVM (Linear kernel) performance metrics	15
6	SVM (RBF kernel) performance metrics	16

List of Figures

Figure	Title	Page
1	(Yu et al.) Univariate COX regression analysis	8
2	Symptom Detection	11
3	Information Extraction from Patient History	11
4	UMAP Projections	12
5	Logistic Regression Plots	13
6	Decision Tree Plots	14
7	SVM (Linear kernel) plots	15
8	SVM (RBF Kernel) plots	16

List of Abbreviations

Abbreviation	Description
ACLF	Acute-on-Chronic Liver Failure
CLD	Chronic Liver Disease
EMR	Electronic Medical Record
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
ILBS	Institute of Liver and Biliary Sciences
MELD	Model of End-stage Liver Disease
NAFLD	Non-alcoholic Fatty Liver Disease
NASH	Non-alcoholic steatohepatitis

Chapter 1: Introduction

Chronic Liver Disease (CLD) is the continuous worsening of liver functions which lasts more than 6-7 months [1]. In this process, there is a generation of harmful proteins and clotting factors, inflammation of liver parenchyma that results in cirrhosis and fibrosis. In some cases, bile is excreted, and detoxification of dangerous products of the metabolism is also observed. Cirrhosis is considered the last phase of the prognosis because it causes changes to the liver structure and triggers neoangiogenesis and nodule formation.

Around 50% of patients with Chronic Liver Failure develop life-threatening conditions like multiple organ failure as the illness progresses. Due to the high mortality rate observed worldwide, there are many scoring systems which use clinical parameters at the time of diagnosis like MELD (Model of End stage Liver Disease) [2], CTP (Child Turcotte Pugh) [3], and the European Association's CLIF-C ACLF [4]. MELD Score is widely used to rank the patients for liver transplant and is calculated using [5]:

- **Serum Sodium**: Measures the fluid-balance regulation in the body.
- **Bilirubin**: Bile is passed through the liver and finally excreted. Measuring the level of bilirubin in blood will tell us how effectively it is being cleared by the liver.
- **Creatinine**: Quantifies kidney function.
- **Internal Normalized Score (INS)**: Indicative of liver functioning and measures the formation of proteins that leads to blood clotting.

In most cases, it has been observed that we get a confirmed prognosis between the third and seventh day after hospitalisation and can plan for liver transplant based on the clinical scores (like MELD).
With the rise in machine learning applications in healthcare, we can filter out the best clinical parameters to predict prognosis. Trained models can be deployed to get predictions at different stages of treatment and automate the decision-making in a clinical setting. Currently, few models exist for liver failure, and even fewer have been tested on Indian patients. This project aims to understand the existing models and develop one using data obtained from Indian hospitals. Effective disease management strategies and early diagnosis can significantly lower the mortality rate.

Objectives

- Using longitudinal clinical data on liver disease patients with some of them having Acute liver failure to predict CLD.
- Making a parser that extracts useful information from EMRs and stores it in a tabular format.

Chapter 2: Literature Review

The dynamic prediction model for the prognosis of ACLF given by Yu et al.[6] is one of the most recent approaches to predicting live failure. These worked with time-series data to evaluate whether changes occurring in the features over time were connected to ACLF prognosis. These temporal evolutions were analysed using Univariate and Multivariate Cox Regression. Patient data used by them was collected over four weeks. Their analysis showed that variations in bacterial infections, age, WGO, gastrointestinal bleeding, etc. were related to ACLF. Their key results are summarised in the following diagram.

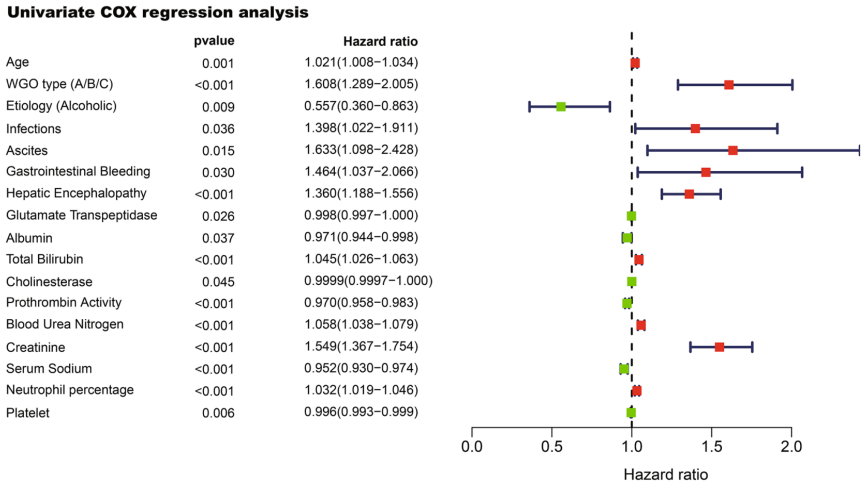


Figure 1: *(Yu et al.) Univariate COX regression analysis*

There are several other papers where the different factors causing chronic liver diseases have been studied.

Another paper which is pertinent to our work assessed the changes in the prevalence of the causes of CLD from 1988 to 2008 [7]. According to them, there has been an increase in the number of CLD cases in these twenty years. Over time, the CLD cases due to Non-alcoholic Fatty Liver Disease (NAFLD) were drastically increasing with the increase in obesity. This major shift led to a decrease in the percentage of cases caused by HCV infections. NASH (Non-Alcoholic SteatoHepatitis) is the most severe type of NAFLD and often evolves into cirrhosis. NASH is very common in diabetic and obese patients and is the reason behind the high mortality rate.

Chapter 3: Methods

Data Description

The data used in this study was collected at the Institute of Liver and Biliary Sciences (ILBS) in New Delhi. All the patients had presented with long-term liver problems, but all the patients did not have CLD. We received the data in two forms, the first was an excel sheet containing 3747 patients and 41 features, but complete information about the diagnosis was given only for 1104 patients. The features given for each patient include physical parameters like sex, gender and age, and multiple clinical parameters obtained after performing blood tests.
Nearly 1600 health records were provided in the form of word documents. Some missing information about the remaining patients can be extracted from them.
In some cases, multiple entries are also present for a patient. We decided to use them as augmentations to increase the size of our dataset.
790 out of 1104 patients had a confirmed diagnosis of CLD, and even the type of CLD was mentioned. These subcategories have been formed based on the cause of CLD. The distribution of patients is as follows:

Type of CLD	**Number of Patients**
CLD Ethanol	459
CLD NASH	159
CLD HBV	41
CLD HCV	37
CLD Cryptogenic	41

Table 1: *Dataset diagnosis distribution*

Types of CLD

1) **CLD Ethanol**: Excessive drinking and alcohol consumption for several years leads to this condition. It is the most common cause of CLD.
2) **CLD NASH**: This is the most severe type of NAFLD. Patients' lifestyle choices are responsible to a large extent as this affects the obese. Controlling calorie and fat intake and leading an active life can reduce the chances of developing NASH. The symptoms of NASH appear once the liver has deteriorated beyond repair [8].
3) **CLD HBV**: CLD is caused by Hepatitis B viral infection.

4) **CLD HCV**: When CLD is caused by hepatitis C viral infection (hepatotropic RNA virus). Nearly 50% of cases of Hepatitis C infections lead to long-term liver problems.
5) **CLD Cryptogenic**: CLD causes irreparable changes to the liver structure and presents in the form of cirrhosis. The reason behind the CLD is unknown.

Preprocessing the Dataset

While cleaning the data, we removed columns where more than 15% of the values were present. This was followed by imputing missing values for the remaining columns with the mean. After this, we dropped the rows of patients with unknown diagnoses.

The final dataset contained 1104 patients and 26 features. Description of the features present in the preprocessed dataset [9][10]:

Clinical Parameters	**Description**
Clotting time (CT)	Time after which clot formation starts after adding the start reagent to the blood.
Clot Formation time (CFT)	Time taken for the clot firmness to reach 20mm after CT.
A5, A10, A15, A20, A25, A30 values	Clot amplitude (firmness) after 5,10,15,20,25,30 mins.
Maximum Clot Firmness (MCF)	Largest observed amplitude.
Alpha-angle	Tangential angle 0 and the curve when clot firmness has reached 20mm.
Maximum Lysis (ML)	Percent of clot stability lost wrt MCF at the end of the test.
Lysis Index after 30 mins (LI 30)	Clot stability wrt MCF (%) is measured thirty minutes after clotting time.
MaxV	Maximum velocity of clot formation

Table 2: *Dataset after preprocessing, descriptions of parameters*

Parser to Extract Information from EMRs

The medical text provided has patient history, treatment details, and test results in paragraph format. This needs to be made tabular before analysis can be performed on it.

I also worked on a parser to extract useful information from these EMRs. It uses the famous biomedical NLP library, SciSpacy [11], to process medical texts. SciSpacy has multiple pre-trained models for entity detection and identifying dependency relations in EMRs. I applied the **en_ner_bc5cdr_md** model, which can segment diseases, symptoms and chemicals appearing in the text. This model has been trained on the BC5CDR corpus. It can be used to create the final diagnosis column of the patient.
Some examples of the results obtained after applying these models are given below.

- Symptom detection:

Chief Complaint

Fever since past since past 15 days

Non productive cough since past 15 days

Altered sensorium since past 2 days

```
{'bc5cdr': {'Fever': 'DISEASE'}, 'craft': {}, 'nlp': {'past 15 days': 'DATE'}}
{'bc5cdr': {'cough': 'DISEASE'}, 'craft': {}, 'nlp': {'past 15 days': 'DATE'}}
{'bc5cdr': {}, 'craft': {}, 'nlp': {'past 2 days': 'DATE'}}
```

Figure 2: Symptom Detection (a): *Example of some symptoms written in the Docx files.*
(b): *The same passed through multiple NLP models to detect those symptoms.*

- Extraction of Information from Patient History

History

Mrs XXX XXXX 50 yrs old female who is a known hypertensive , and had hypothyroidism , non diabetic had an index issue in the form of fever since past 15 days which was high grade , intermittent associated with gernalised body aches (Backaches +) . patient also complaint of non productive cough since past 15 days . later patient also developed bleeding PV since past 2-4 days , fever settled . following that patient again developed cough and also he had episode of altered sensorium since past 2 days. patient was evaluated outside and was found to have dengue NS1 postive and typhoid IgM positive . Patient was admitted and was being managed conservatively with RDPC, iv fluids and other supportive measures . patient came to ILBS with above mentioned complaints for further evaluation and management . There is no , vomiting, abdominal pain, altered bowel habits, hematemesis, and malena, burning micturition or decreased urine output. There is no h/o any intoxications, indigenous medications, major surgeries, blood transfusions or IV drug abuse prior to onset of the disease. There is no h/o CAD/TB/COPD.

```
{'bc5cdr': {'hypertensive': 'DISEASE', 'hypothyroidism': 'DISEASE', 'diabetic': 'DISEASE', 'fever': 'DISEASE', 'a
ches': 'DISEASE', 'cough': 'DISEASE', 'bleeding': 'DISEASE', 'dengue': 'DISEASE', 'typhoid': 'DISEASE', 'vomiting
': 'DISEASE', 'abdominal pain': 'DISEASE', 'hematemesis': 'DISEASE', 'malena': 'DISEASE', 'drug abuse': 'DISEASE'
}, 'craft': {'PV': 'GGP', 'drug': 'CHEBI'}, 'nlp': {'50 yrs old': 'DATE', 'past 15 days  ': 'DATE', 'past 15 days
': 'DATE', 'PV': 'ORG', '2-4 days': 'DATE', 'past 2 days': 'DATE', 'RDPC': 'ORG', 'malena': 'GPE', 'CAD/TB/COPD':
 'ORG'}}
```

Figure 3: Information Extraction from Patient History
(a): *Example of patient history written in the Docx files.*
(b): *The same passed through multiple NLP models to extract information about the history like previous diseases and basic information about the patient.*

Chapter 4: Results

Projection and Clustering

I have used UMAP to project the high-dimensional data (with 26 features) onto a 2D plot to see whether there is any rough similarity between different groups. UMAP has recently been shown to be consistently better and more stable than other projection methods like t-SNE.
However, from the projection plot with the actual labels, we can see that there is little separation between the two classes (CLD-Ethanol and CLD-non Ethanol). So, we can expect the supervised results not to do too well either, as the data is non-clusterable.
I also do an unsupervised K-Means clustering of the data and show the results on the same projection.

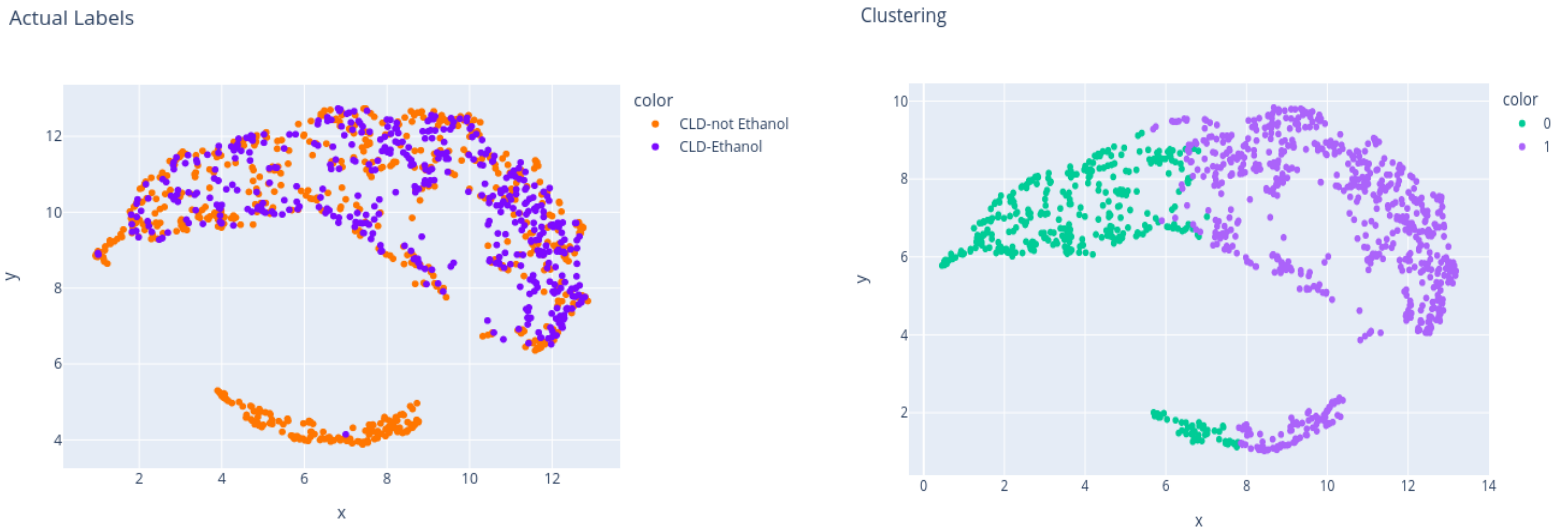


Figure 4: *UMAP projection coloured with* **(a):** *actual labels* **(b):** *K-Means clustering labels.*

We trained different types of models on the preprocessed dataset and evaluated their performance using multiple metrics. The results for logistic regression, decision trees and support vector machines (SVM) can be found below. 10-fold stratified cross-validation was performed to reduce bias that might occur from train-test random splitting.

Logistic Regression

Training Accuracy	67.381
Testing Accuracy	63.578
Precision	0.559
Recall	0.588
F-score	0.573
AUC-ROC	0.629

Table 3: *Logistic Regression performance metrics*

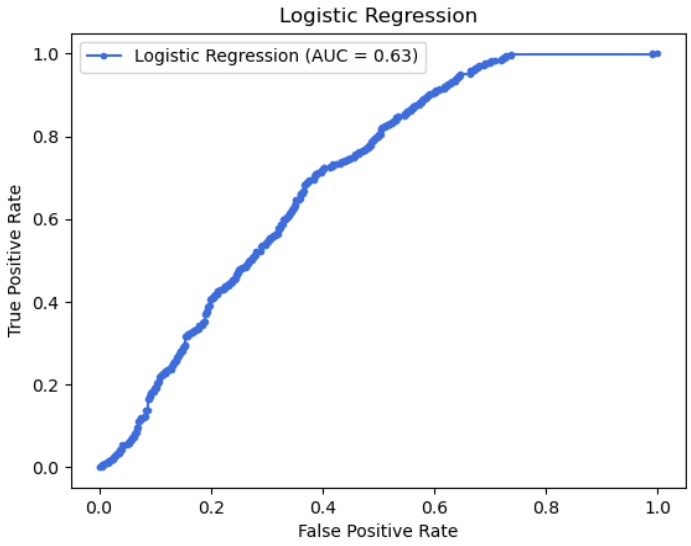
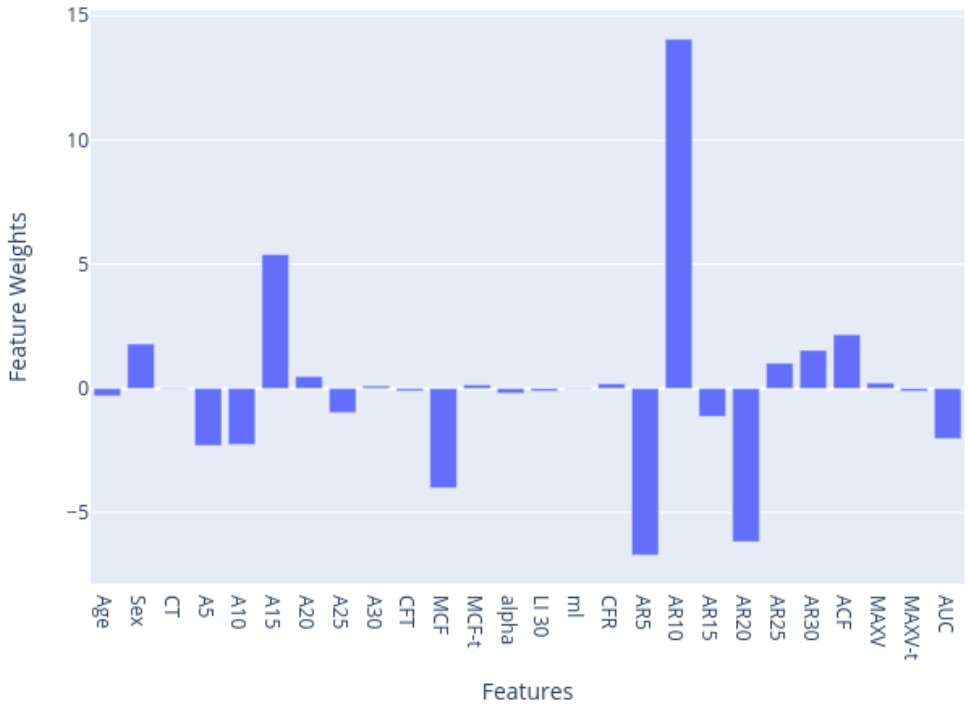


Figure 4: Logistic Regression Plots
(a): *AUC-ROC curve*
(b): *Feature weights*

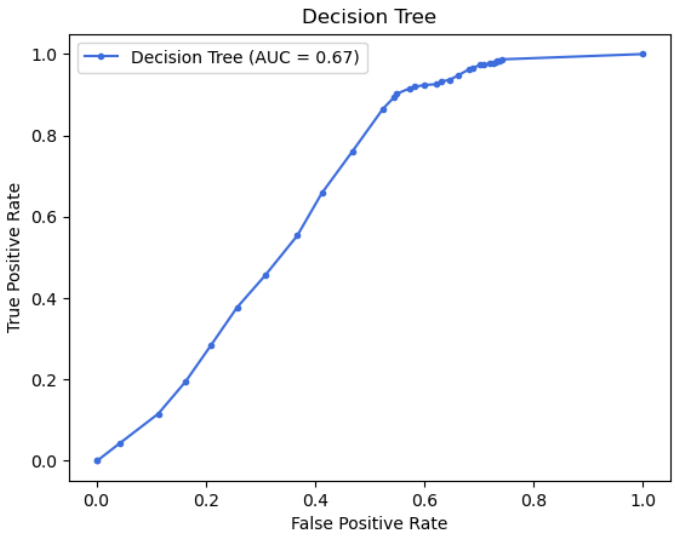


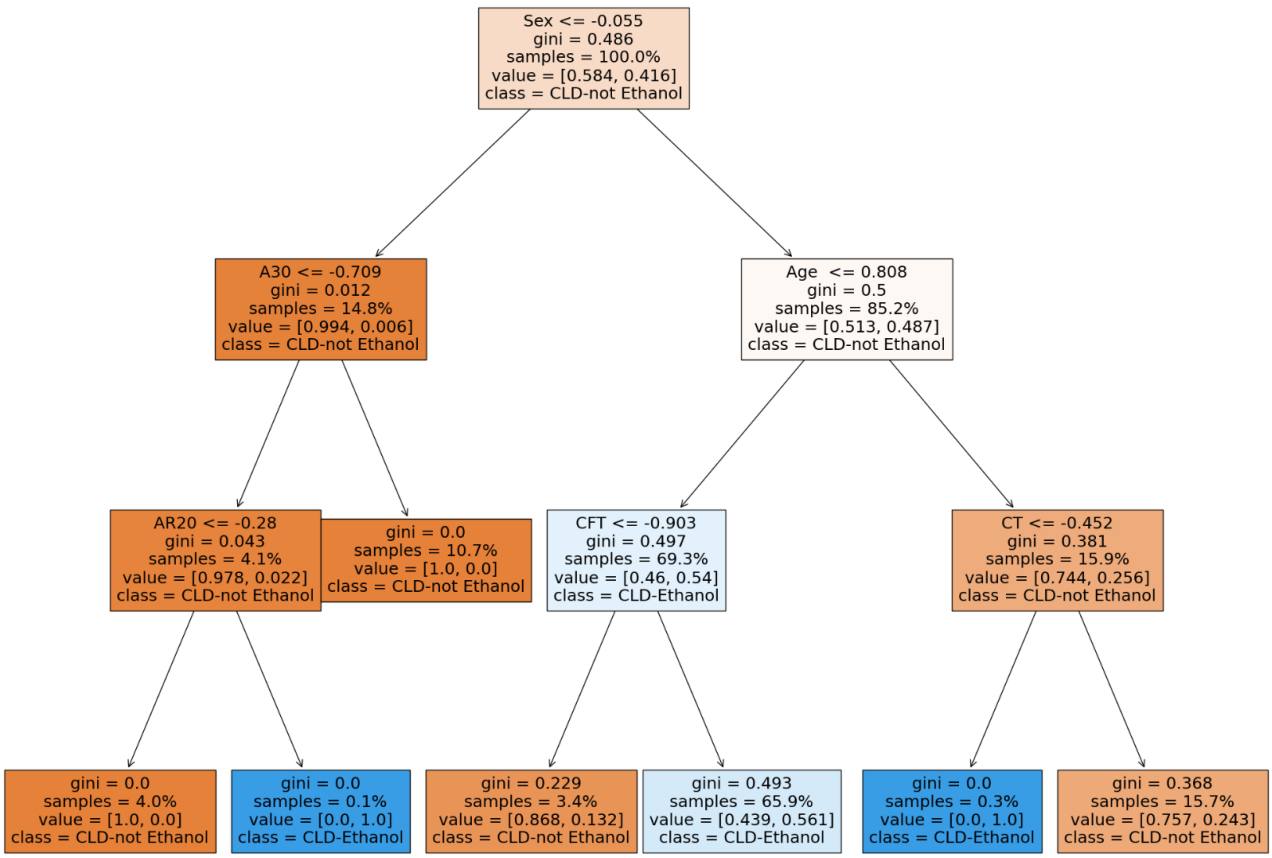
Decision Tree

Training Accuracy	66.576
Testing Accuracy	63.758
Precision	0.540
Recall	0.865
F-score	0.665
AUC-ROC	0.670

Table 4: *Decision Tree performance metrics*

Figure 5: Decision Tree
(a): *AUC-ROC curve*
(b): *The decision tree obtained*



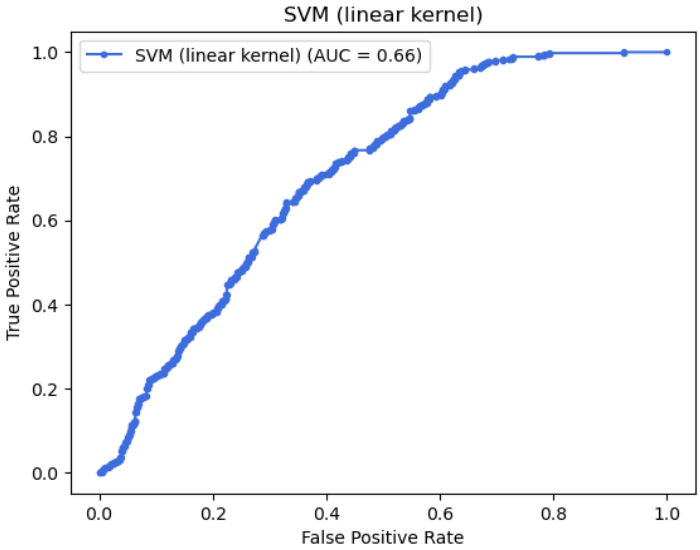


Support Vector Machine

a) Linear Kernel

Training Accuracy	66.938
Testing Accuracy	65.659
Precision	0.573
Recall	0.682
F-score	0.623
AUC-ROC	0.660

Table 5: *SVM (Linear kernel) performance metrics*



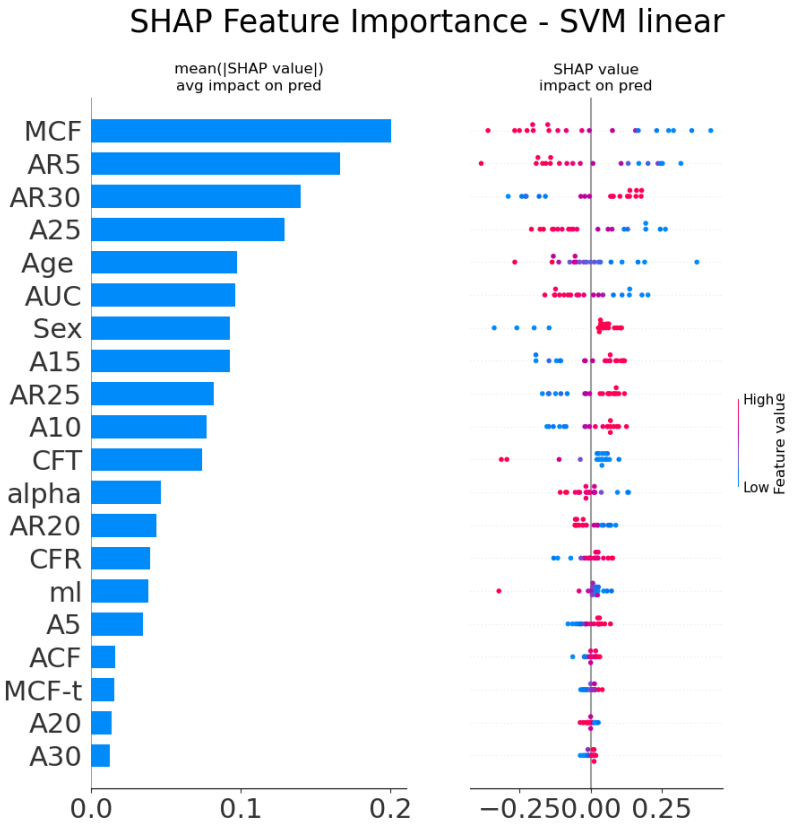
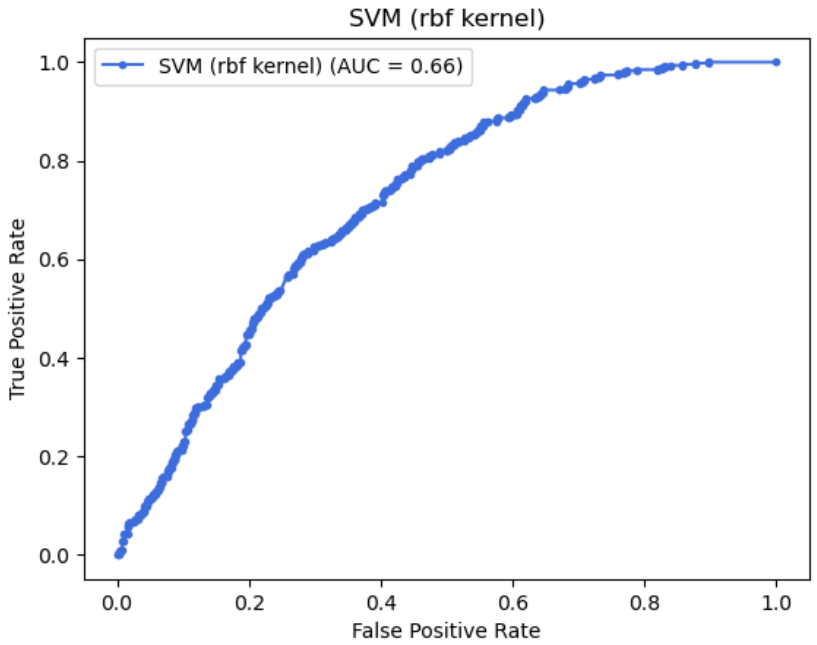


Figure 6: SVM (Linear Kernel) plots
(a): *AUC-ROC curve*
(b): *SHAP feature importance plots*

b) RBF Kernel

Training Accuracy	70.853
Testing Accuracy	66.567
Precision	0.593
Recall	0.623
F-score	0.608
AUC-ROC	0.660

Table 6: *SVM (Radial Basis Function / RBF kernel) performance metrics*



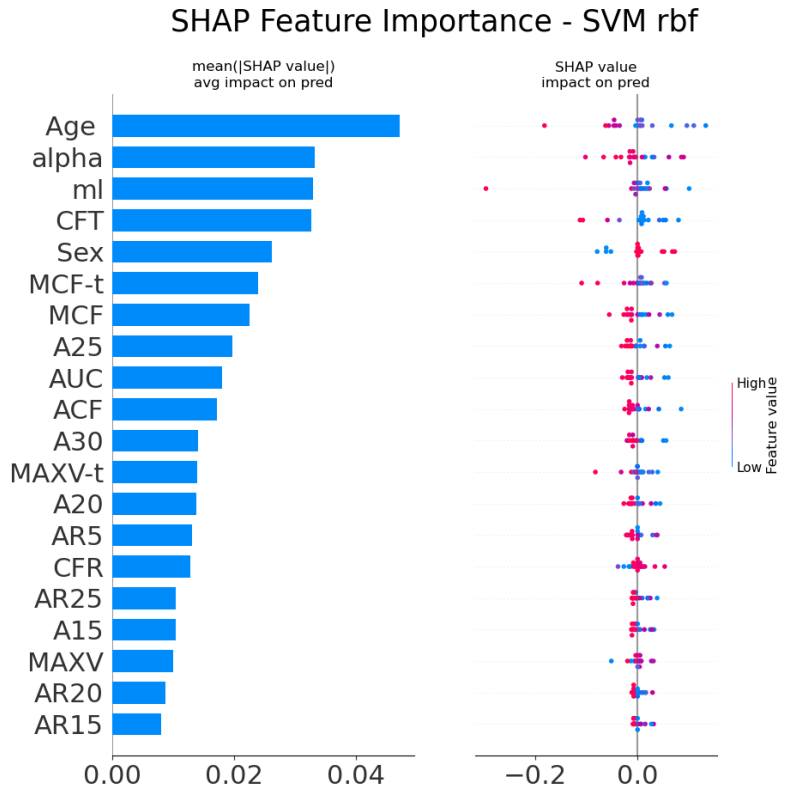


Figure 6: SVM (RBF Kernel plots)
(a): *AUC-ROC curve*
(b): *SHAP feature importance plots*

Chapter 5: Conclusion

The analysis we have performed so far has given us the most important of the parameters that predict whether a patient has CLD-Ethanol or some other type of CLD. These include factors like age, sex, AR5-30 values (clot amplitude), ML (maximum lysis), alpha (angle), CFT (clot formation time (20mm)), CT (clotting time) etc. from the SHAP feature importance plots and the splitting nodes of the decision tree. The best performance in terms of training and testing accuracy was achieved by

We can also observe that the AUC of all these models is not very high. This can be explained to a great degree by the largely overlapping nature of the data (as seen in the UMAP projections), so it would not be possible to separate those points with an explainable/simple hyperplane.
This classification can be very useful in determining the medication for patients who do not want to share/lie about their drinking habits and alcoholism, as most of the time, the medication for diseases is different as the general medication would cause harm to the said alcoholic patient.

Next Steps

We want to apply the parser built using Scispacy and Med7 [12] on all the patient files available and extract the features provided in the excel sheet. These patients can now be used to test the performance of the trained models. We would get useful insights into the stability of our model. If the results are not as good as expected, we can re-train them with more data and evaluate them again. Currently, multiclass classification for all the CLD types was not possible because the cases due to HBV and HCV were negligible (lesser than 4%). If there are sufficient patients for these categories in the files, then we can extend this to a multiclass classification problem.
We also want to try to classify bleeders vs non-bleeders based on these parameters.

References

1. Website. Available: Sharma A, Nagalli S. Chronic Liver Disease. [Updated 2022 Jul 4]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554597/

2. Kamath PS, Kim WR. The model for end-stage liver disease (MELD). Hepatology. 2007;45. doi:10.1002/hep.21563

3. Tsoris A, Marlar CA. Use Of The Child Pugh Score In Liver Disease. StatPearls [Internet]. StatPearls Publishing; 2022.

4. Engelmann C, Thomsen KL, Zakeri N, Sheikh M, Agarwal B, Jalan R, et al. Validation of CLIF-C ACLF score to define a threshold for futility of intensive care support for patients with acute-on-chronic liver failure. Crit Care. 2018;22: 1–8.

5. Understanding MELD Score for Liver Transplant. In: UPMC | Life Changing Medicine [Internet]. [cited 23 Sep 2022]. Available: https://www.upmc.com/services/transplant/liver/process/waiting-list/meld-score

6. Yu Z, Zhang Y, Cao Y, Xu M, You S, Chen Y, et al. A dynamic prediction model for prognosis of acute-on-chronic liver failure based on the trend of clinical indicators. Sci Rep. 2021;11: 1–13.

7. Younossi ZM, Stepanova M, Afendy M, Fang Y, Younossi Y, Mir H, et al. Changes in the prevalence of the most common causes of chronic liver diseases in the United States from 1988 to 2008. Clin Gastroenterol Hepatol. 2011;9. doi:10.1016/j.cgh.2011.03.020

8. Nonalcoholic fatty liver disease. In: Mayo Clinic [Internet]. 22 Sep 2021 [cited 23 Sep 2022]. Available: https://www.mayoclinic.org/diseases-conditions/nonalcoholic-fatty-liver-disease/symptoms-causes/syc-20354567

9. Crochemore T, de Toledo Piza FM, dos Reis Rodrigues R, de Campos Guerra JC, Ferraz LJR, Corrêa TD. A new era of thromboelastometry. Einstein . 2017;15: 380.

10. Thromboelastometry. Wikimedia Foundation, Inc.; 15 May 2009 [cited 23 Sep 2022]. Available: https://en.wikipedia.org/wiki/Thromboelastometry

11. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. 2019 [cited 23 Sep 2022]. Available: https://www.aclweb.org/anthology/W19-5034.pdf

12. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A. Med7: A transferable clinical natural language processing model for electronic health records. Artif Intell Med. 2021;118. doi:10.1016/j.artmed.2021.102086