IS 733

Final Project Report

December 11, 2020

Predicting Hepatitis C Virus (HCV) using Machine Learning Model

Saydeh Karabatis, Masnoon Nafees, Sabrina Mamtaz Nourin, Argho Sarkar,Simran Shaik
{saydeh1, masnoon1, snourin1, asarkar2, simran2}@umbc.edu

Abstract

In this work we address the problem of diagnosing Hepatitis C Virus (HCV) its impact on the liver, and identifying healthy patients who could possibly donate a portion of their liver to patients with advanced stages of Hepatitis C. The process of diagnosing HCV includes physical examination, imaging, and blood analysis. If the results are not clear, the diagnosis is made based on pure guessing which is not reliable. Therefore, we implemented a 2-phase machine learning prediction algorithm to assist the medical professionals in detecting the HCV. We downloaded a publicly available dataset of HCV patients and we addressed the problem of both missing data and highly imbalanced classes. We applied supervised and unsupervised learning techniques to the dataset and the results were quite impressive. We have achieved an accuracy of 98%, which is much higher than the published results on the same dataset.

1 Introduction

If one would try to define the human body in a technical term, one could think of a machine learning process in which certain parts of the body work together, each taking control of a sub-process, then integrating all together to generate a model: the human body. An important feature of machine learning is data pre-processing; it takes raw data, removes the unnecessary values, and regenerates clean data ready for analytical use. Likewise is the human body, it continuously preprocesses things, whether they may be thoughts in the mind or material in the body. Certain types of human-related pre-processing tasks are achieved through toxin filtering stations physically located inside our bodies. These stations are responsible for taking in material, breaking it apart, generating useful substances, and throwing the toxins away. If toxins are not removed, poison builds up in the body and creates brain malfunction leading to Encephalopathy. The liver, a participant organ in the food digestion process, is an excellent example of a toxin filtering station. It detoxifies the blood and reproduces healthy red blood cells. However, the liver can get inflamed and damaged by certain diseases. Contaminated food, saliva transferred from one body into another, needle sharing, excessive alcohol drinking, and blood transfusion are examples of activities that could cause inflammation of the liver.

Inflammation is an enemy of the human body by all means. “Hepatitis means inflammation of the liver” is caused by a virus in most cases [3]. The most common types of hepatitis are Alcoholic Hepatitis, Hepatitis A, B, and C. Alcoholic Hepatitis is caused by excessive drinking of alcohol and leads to severe scarring of the liver. Hepatitis A is caused by contaminated food or water consumption. Hepatitis B “is passed from person to person through blood, semen, or other body fluid”, and finally type C is caused by a virus called Hepatitis C (HCV) [32]. It widely known that HCV is contracted during activities such as sharing contaminated syringes or through infected-blood transfusion.

According to the US Department of Veteran Affairs, Hepatitis C Virus is the most common blood-borne disease in the United States [32]. Some HCV infected people exhibit short-term symptoms, while others carry the symptoms for a longer period of time. The damage of the liver caused by HCV can be mild, moderate, or severe resulting in life threatening health problems. When in early stages, the liver condition is less severe and the patient can be referred to as “Hepatitis C” patient. Medical treatment and proper diet can be successful at this point provided that the patient will not contract the virus again [3]. When the infection level is in the middle stage, the patient is known to be having “Fibrosis”, but when the patient is an advanced and final stage, the liver is experiencing “Cirrhosis” having the least chances of recovery and possible liver cancer/death. The World Health Organization (WHO) states that it takes between 20-25 years for the HCV-infected liver to reach the Cirrhosis stage and an additional 5-10 years to develop into cancer and death as shown in Figure 1 [26].

Preventing Hepatitis C infection requires avoiding blood and needle sharing. Given that drug addiction has risen over

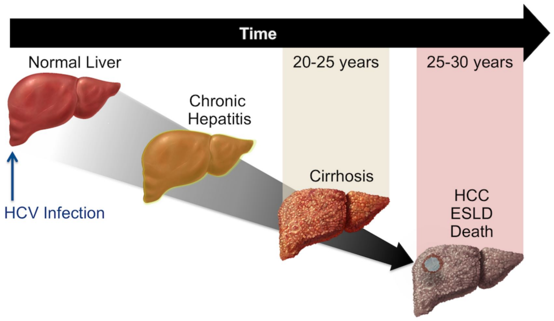
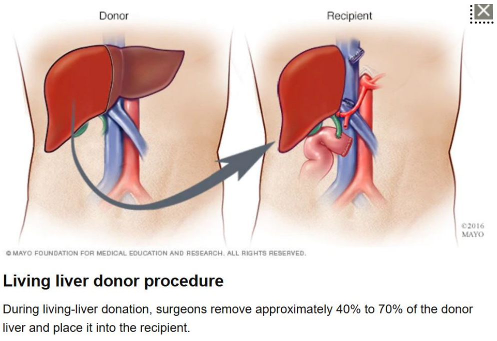


Figure 1: Time Progression of Hepatitis C


Donor
Recipient
©2016
MAYO
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Living liver donor procedure

During living-liver donation, surgeons remove approximately 40% to 70% of the donor liver and place it into the recipient.

Figure 2: Taken from the Mayo Clinic [5]

the years, so are the number of HCV cases. However, early medical intervention can lead to a victory over the virus. "Getting tested for hepatitis C is important, because treatments can cure most people between 8 to 12 weeks"[3]. An HCV diagnosis requires physical examination, general laboratory studies, and hepatic inflammation function test of the blood[25]. Based on the results, clinical experts can apply a set of rules, decide on the fate of the liver, and a treatment can be administered to the patient. When medications fail to succeed, an HCV-infected patient could become a potential recipient of a donated portion from a healthy liver. For the liver transplant procedure to take place, it requires a donor willing to donate a portion of her/his liver, a match between the recipient and the donor, two invasive surgeries to both donor and recipient, and the ability of the recipient's body to accept the transplanted liver as shown in Figure 2. However, in some cases, the recipient's body could recognize the transplanted liver as a foreign object and attack it. The transplant process fails and death in inevitable.

The above diagnostic approach using set of rules has been widely used and has been leading to good outcomes. However, success is not 100% guaranteed due to the fact that it fails to identify some HCV-infected patients who do not exhibit all the symptoms. To address this problem, one can take advantage of machine learning algorithms and design a model that will increase the accuracy of the number of diagnosed HCV cases leading to early intervention and better success of the treatment. In this paper, we propose to create such a model to assist the health care providers have better success rates in the diagnosis that will lead to better treatment results. To achieve this goal, we downloaded a labeled HCV dataset from the UCI ML Repository [18]. The dataset contains results of blood tests done on patients who have been infected with the HCV virus mixed with other healthy blood donors. The blood donor samples have been tested and no HCV has been found in any. The HCV-infected patients can exhibit any of the three stages of the liver.

The main purpose of the study is to implement a 2-phase machine learning prediction algorithm and to assist the medical professional in detecting HCV. Our main contribution is applying supervised and unsupervised learning techniques to the blood tests results in order to:

- Identify HCV infected subjects and potential liver donors
- Identify stages of HCV infected liver and determine whether a medical treatment or liver transplant is required

The remaining of the paper is as follows. Section 2 provides an overview of related work in this area and a comparison to our method. Section 3 outlines a detailed description of the dataset. Section 4 explain the steps of the data preprocessing Section 5 justifies the performed methodologies and describes the unsupervised and supervised analytical algorithms used in the study. Section 6 presents evaluation and accuracy results, and finally we reflect on future work in Section 7.

2 Related Work

Hepatitis C was first isolated from other hepatitis diseases in 1989 by Choo [23]. Infectious disease specialists identified certain markers in the human body to help detect the virus based on three scales: Below Range, Normal Range, &

Above Range as described in [6, 7, 22]. Since then, it is estimated that 170 million people world wide carry the virus, or 3% of the global population. Due to the fact that we, the authors of this paper, have no medical background and to better understand liver diseases, we relied on trusted medical resources such as Mayo Clinic, WHO, US Department of Veteran Affairs, and the CDC to gain some initial understanding needed for this project [5, 26, 32, 3, 25]. The knowledge provided by these resources was very beneficial in learning facts about HCV that helped shape the project.

Recent studies have shown that one can take advantage of machine learning algorithms to identify and confirm the different stages of the HCV-infected liver [19]. Hoffman et al. used a published study on hepatitis C patients where they provided algorithms that produce plausible decision trees [28] which detect liver fibrosis and cirrhosis. Validation of the decision trees using the leave-one-out method resulted in a model that had an accuracy close to 80% which was higher than the baseline of pure guessing. An interesting aspect of this work is that they use the same published dataset that we also use in this research where our algorithms resulted in an accuracy of 98% using J48 [30] and PART [9].

Another related work on Hepatitis C and machine learning is from [8]. The authors used descriptive analysis to identify patients with symptoms of HCV 2–3 years before they were diagnosed. By using Logistic Regression [15], Random forest [2], and gradient boosting trees they were able to achieve 95% precision for all algorithms. In comparison with our models we were able to achieve precision of 98% by only using J48.

Published results in the Online Journal of Public Health Informatics show that by using the Random Forest ensemble classifier, one can highly predict HCV with recall close to 99% and precision $\bar{8}9\%$ [14]. The dataset used in their study contains about 100,000 instances with highly imbalanced classes. There is no information about the input variables. In comparison to our approach using Random Forest, we were able to reach 97% precision and 78% recall.

To carry out the study, we downloaded a dataset from the UCI Machine Learning repository [18]. This dataset was compiled by a lab in Germany and included blood test results from 516 patients. We applied most of the Data Mining and Machine Learning techniques found in the "Data Mining" and "The WEKA data mining software" books [33, 10].

3 Description of the Dataset

To conduct the research, we downloaded a labeled **"HCV Data Set"** from the UCI Machine Learning Repository life domain field [18]. The data set contains demographic information and laboratory results of blood tests collected from a lab in Germany. The data set has 13 input variable, one class, and 615 instances, each having multi-variate characteristics. The attribute values of the instances are categorical and numeric data types. Each patient is given a unique identifier and is represented under the ID column of the dataset.

Demographic attributes include age and gender:

- **Age** of patient in years at time of the blood test and is of integer data type. The age of the patients falls between 19 to 77 years old. Most of the subjects fall between 35 to 60 years of age.
- **Sex** signifies the reported gender of the patient and is of categorical value, female and male. The ratio of females to males equals 2:3.

Since the data set was created by a laboratory in Germany, the values of the attributes are specific to the lab. Blood lab results include results of metabolic and liver analysis. We list the name of each with brief description of its importance to the liver:

- **ALB**: Albumin, is a protein made by the liver and helps keep fluid in the bloodstream so it doesn't leak into other tissues of the body. The unit measure in gram/litre(g/L).
- **ALP**: Alkaline Phosphatase, an enzyme found in the body and could show signs of liver or bone disease. The unit measure is in units/litre (U/L).
- **ALT**: Alanine Aminotransferase, an enzyme found mostly in the liver and kidneys. It is normally found to have very low levels in normal blood tests, but when the liver is unhealthy, these levels increase in the blood. The unit measure is units/litre (U/L).
- **AST**: Asparate Aminotransferase, an enzyme found mostly in the liver and heart. It is normally found to have very low levels in normal blood tests, but when the liver is unhealthy, these levels increase in the blood. The unit measure is units/litre (U/L).

- **BIL**: Bilirubin, an orange-yellow pigment found in the red blood cells. Normal to high levels could indicate liver issues. The unit measure is micromole/litre (µmol/L).
- **CHE**: Choline Esterase, an enzyme found in the liver, pancreas, and other parts of the body. It is an excellent bio marker for find out of the liver has cirrhosis. The unit measure is kilo units/litre (KU/L).
- **CHOL**: Cholesterol, a fatty substance found in all the cells of the body and is decomposed into subcategories. The unit measure is micromole/litre (µmol/L).
- **CREA**: Creatinine, a waste product produced by the muscles and is disposed by the kidneys. The results could reflect the functioning levels of the kidneys. The unit measure is micromole/litre (µmol/L).
- **GGT**:Gamma Glutamyl Transferase, an enzyme found in many organs in the body, but it is mostly present in the liver. Elevated values of GGT could indicate liver malfunction. The unit measure is units/litre (U/L).
- **PROT**: Protein, an measure that indicates the different protein levels in the body. Low levels of protein could indicate liver problems. The unit measure is gram/litre(g/L).

The class contains 5 values: "Blood Donor", "Suspect Blood Donor", "Hepatitis C", "Fibrosis", and "Cirrhosis". The first class "Blood Donor", belongs to one category, name it One, and signifies blood being free of HCV, while the last 3 classes "Hepatitis C", "Fibrosis", and "Cirrhosis" belong to another category, name it Two, and signify liver disease. The ratio of Category One to Category Two is about 6:1, whereas the ratio of Category One to each class in Category Two is about 17:1 signifying the presence of highly imbalance classes. The last class "Suspect Blood Donor" was not tested and suspected to be clean from HCV. We treat this class differently as it will naturally belong to the testing data. We will illustrate more on this in the data pre-processing section. Detail description of each class is as follows:

- **Blood Donor**: Blood lab results from patients who have healthy liver
- **Suspect Blood Donor**: Blood lab results from patients who might have have healthy liver
- **Hepatitis C**: Patients having an early stage of Hepatitis C
- **Fibrosis**: Patients having middle stage of Hepatitis C who could be candidates for liver transplant
- **Cirrhosis**: Patients having advanced stage of Hepatitis C who could be candidates for liver transplant.

4 Pre-Processing the Data

The purpose of data preprocessing is to prepare the data for analysis and model creation. Data could be cleansed from **outliers**, **missing values** should be replaced with meaningful values, values could require **conversion**, some input variables need to be placed in **bins**, continuous values could be **normalized**, features relevant to the output should be **selected**, and **new features** could be constructed if needed [16].

- **Outlier Detection**: Data outliers are data points that are spread away from the rest of the data points. Outliers do not always suggest that these values need to be eliminated. In some cases, outliers mean a high spike in a level, in other times, they could be due to data entry errors. Understanding the data will help decide if outliers are useful and should be kept or removed. In case of the HCV dataset, there are no existing outliers. This was also asserted from the authors of the dataset through an email exchange.
- **Missing Values**: Having missing values in a real life science data is unavoidable. The HCV dataset was generated from real patients who provided their results of blood tests for research. The missing values are due to the fact that not all the patients had their blood examined for each feature. We discovered 31 missing values in the dataset and had to make some choices:
  - Ignore the missing values. The HCV dataset is relatively small. Deleting raws not only reduces the number of instances and could be problematic in our study.
  - Impute the missing values with the mean value of every feature respectively. Example: if values are missing in the ALT column, using the unsupervised ReplaceMissingValues() filter in weka [10], we replace these values with the mean value of all existing results of ALT.
  - Predict the missing values by creating a classifier for each column and replace each missing value by the predicted values.

Examining the data thoroughly made us realize that choosing only one approach might not satisfy the study. Instead, we decided to apply the first two approaches and generate two datasets for each approach. Dataset D1 has no missing values, while dataset D2 contains imputed missing values. Due to time limitations we could not explore option 3 and would like to do so in future work.

- **Data Conversion**: To make sense of the data, values need to be compared to some existing standard values. The HCV dataset was created in a German laboratory that uses different units than the International System of Units (SI Units). For the purpose of understanding the values of the data, a mapping tool needs to be implemented that utilizes unit conversion as seen in the conversion table. Table 1 was constructed based on information collected from Quest Diagnostics(QD), Lab Corp, and NIH websites [22] [7] [6]. We preserved the existing values and created a new column for each attribute that requires conversion into SI Unites, using the formula: *newcolumns= oldcolumn * conversion coefficient*

Blood Test Code	UCI Unit	Quest Diagnostic Unit	Molar Mass	Convert to QD Numbers	Coefficient
ALB	g/l	g/dl		times 10: 1L = 10 dL	0.1
ALP	u/l	u/l		times 1	1
ALT	u/l	u/l		times 1	1
AST	u/l	u/l		times 1	1
BIL	umol/l	mg/dl	584.7 g/mol	multiply by 0.0585	0.0585
CHE	ku/l	iu/dl		times 1000: 1KU = 1000IU	1000
CHOL	umol/l	mg/dl	386.7 yg/mol	multiply by 38.61	38.61
CREA	umol/l	mg/dl	113.12 g/mol	multiply by 0.0113	0.0113
GGT	u/l	u/l		times 1	1
PROT	g/l	g/dl		times 10: iL = 10dL	10

Table 1: Conversion Table

- **Binning**: In case of continuous values, binning (discretization) reduces the number of values to smaller values. Since large number of possible values could contribute to the ineffectiveness of the Machine Learning process, binning is used to store a range of numbers into one bin [16]. According to Quest Diagnostics, the levels of the blood analysis are divided into ranges based on age and gender [7]. To perform binning, we stored the dataset in a relational database server, used sql statements to manipulate it, and to create new binned columns. The values of the newly binned columns are grouped by "Below Range", "In Range", and "Above Range" according to the recommendation of Quest Diagnostics and LabCorp.
- **Data Normalization**: When features often exhibit large range of values, scaling them down will reduce the difference between the minimum and the maximum values of these features. The HCV dataset does not include values with large data range; therefore, no data normalization is required.
- **Feature Selection**: This technique is used to eliminate attributes that are not relevant to the outcome and to emphasize some that are useful to predict the outcome of the model. Using various tools we were able to calculate data impurity by applying the *gain ratio* and the *variable importance* methods, then generating a *gini index tree* to confirm our choices. Based on these methods, we were able to identify that AST, ALT, & GGT had the highest predictive abilities compared to the remaining feature. Since the dataset consist of only 14 features, we plan to include the attributes that exhibited lower predictive abilities.

BILL	BILL CU	BILL BIN
7.5	0.43875	In Range
3.9	0.22815	In Range
6.1	0.35865	In Range
18.9	1.1056	In Range
9.6	0.5616	In Range

Table 2: New Feature Construction Sample

- **Feature Construction**: Creating new features “may lead to a more concise and accurate classifiers”[16]. This has been discussed an implemented in both **Data Conversion and Discretization** subsections. To summarize, two columns have been generated: one to convert the data into SI units and the other to create bins for each blood result values, as seen in Table 2.

- **Highly Imbalanced Classes**: As described in the previous section, the classes in the dataset are highly imbalanced. The initial study did not guarantee good accuracy results using the 4 classes and we decided to combine all HCV classes into one class. Such approach generated a more manageable datasets D3 and D4 with ratio 6:1 (Healthy:HCV). D3 is created from D1 (no missing values) and D4 is created from D2 (mean imputed missing values).

- **Newly Generated Datasets**: Since the main purpose of the study is to implement a 2-phase machine learning prediction algorithm that will not only assist the medical professionals in detecting HCV, but will suggest the type of treatment depending on the condition of the liver (medicine or liver transplant), we needed to isolate the HCV patients in one dataset D5 (75 values) with initial class allocations for each patient (Hepatitis, Fibrosis, Cirrhosis). Three new datasets have been generated based on the initial dataset:
  - D3, used in Phase1: No missing Values, 486 instances, 2 classes (HCV or Donor)
  - D4, used in Phase1: Mean imputed missing values: 516 instances, 2 classes (HCV or Donor)
  - D5, used in Phase2: Mean imputed missing values: 75 instances, 3 classes (Hepatitis, Fibrosis, Cirrhosis)

  Now that the data is been pre-processed it is ready for analysis.

5 Methodology

The process of diagnosing HCV includes physical examination, imaging, and blood analysis. If the results are not clear, the diagnosis is made based on pure guessing [19]. HCV is completely treatable if caught at the early stages by administering medicine and following a proper diet [12]. Since it is very important to diagnose HCV as early and as accurately as possible, we intend to find a method that can detect HCV with the highest accuracy possible. We have applied several supervised and unsupervised machine learning algorithms on our HCV dataset. For unsupervised learning, we used Association Rule Mining [17] and Expectation Maximization (EM) [21]. We used the Support Vector Machine [24] method to create boundaries. For supervised models we used: Naive Bayes [29], Logistic Regression [15], Stochastic Gradient Descent [1], K-nearest neighbor [27], Decision Tree [28], and PART [9]. We have also used the ensemble Random Forest [2] and AdaBoost methods[31] . We have evaluated these methods based on accuracy, precision, recall, and by plotting the ROC. We have added brief descriptions of our used methods in the following sections.

5.1 Unsupervised Learning

When data is not labeled, machine learning relies on unsupervised learning to examine the data and look for undetected patterns with minimum support from the human. Following are the brief descriptions of the unsupervised learning methods that we seemed fit to be applied in our dataset:

5.1.1 Discovering Rules - Association Rule Mining

For any particular data set, the main factors that we consider are support, confidence, and lift. Support means counting the number of occurrences of the if {X then Y} relationship in the dataset. Confidence is the number of times these relationships are found to be true. Lift is the rise in probability of having Y with the knowledge of X being present divided by the probability of having Y without any knowledge about X. Apriori algorithm [20] is used in the project for association rules generation.

5.1.2 Discovering Rules - PART (Obtaining rules from partial decision trees)

PART uses the separate and conquer method based on C4.5 to generate a a decision list of rules [30]. During this process, integration of construction and pruning operations needed to be take place to generate a stable tree that cannot be simplified any further. To reduce overfitting in noisy situations,it is necessarily recommended to produce rules that are not even perfect on the training part of the dataset. For larger datasets , this rule functions very well [9]. Though PART is a supervised learning method, we list it in the unsupervised section because we chose to place the rules in one section.

5.1.3 Expectation Maximization (EM)

Expectation Maximization (EM) is popular clustering algorithm and a bit complicated than known K-mean clustering [13]. It uses maximum likelihood to map the density of the data set in order to create the probability of the the distribution. It performs the maximum likelihood estimation on the latent variables and allows the user to control the shape of the cluster. In our project, we used both EM and k-means to generate the clusters.

5.2 Supervised Learning

Supervised learning is a very powerful part of the machine learning process as it maps an input to an output based on a given data. Contrary to the unsupervised learning techniques, the supervised methods use label data to infer an algorithm and to predict the output either through linear regression or through classification. Following are the brief descriptions of the supervised learning methods that we seemed fit to be applied in our dataset:

5.2.1 Naive Bayes

A Naive Bayes classifier is probabilistic and built based on Bayes Theorem. It is commonly used as a base classifier because it assumes that all predictors are independent of each other and all predictors have equal effect on the outcome. Knowing that there is no independence among he input features, we use Naive Bayes as a base classifier.

5.2.2 Discovering Boundaries -Support Vector Machine

Support Vector Machine use linear models to implement non linear class boundaries. It addresses the problems of overfitting and runtime as if in case the cubic coefficients increase for linear model, the training may be infeasible (runtime issue) , resulting model may become too non linear called as over fitting of data as they learn particular linear boundary known as maximum margin hyper-plane is visualized in that instances closer to hyper-plane are called as support vectors , generally support vector machine are linear learned classifiers. We used LibSVM [4] in this study.

5.2.3 Decision Tree

Decision tree is one of the most powerful and popular tool for classification and prediction. A Decision tree is tree structure, which works like a flowchart where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

A tree is learned by splitting the source set into subsets based on an attribute value test. There are several splitting methods of decision trees. In this work, we have experimented with 2 of them - Gini Index and information gain. We have experimented with decision tree with information gain in WEKA [10]. The splitting process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.

When a new instance (with a set of attributes) are tested in a decision tree, the attributes are tested against the values of each node, and assigned to a branch. This testing is repeated until any leaf node is reached. The class in the leaf node represents the class of the instance. In this work, we have assigned the minimum number of leaves per nodes to 2, which indicate that a node will be branched if it has 2 or more instances.

5.2.4 Logistic regression

Logistic regression is used to predict the probability that a given data entry belongs to a certain category. This regression algorithm is very helpful and accurate in binary classifications, but not as great for multiple classes. Logistic regression models the data using the sigmoid function [11], and it assumes that the data follows a linear function. We applied the logistic regression techniques and found it to be comparable to the best classifiers.

5.2.5 K-Nearest Neighbor

KNN is a very common supervised algorithms which can be used to solve both classification and regression problem. Based on the number of nearest neighbour KNN picks the closest one or assumes the similarity in near proximity. The number of neighbors K is always given. For our dataset the KNN is suitable for acquiring results. We tried several number of neighbors (i.e, variable k) and got the different results.

5.2.6 Random Forest

Random forest [2] comprises a large number of individual decision trees that operate as an ensemble. Every single tree in the random forest offers a class prediction and the class with the most votes becomes the prediction of our model. All the individual tree assumed to be uncorrelated. Random forest, unlike decision tree, is less sensitive to data. This is because bagging (sample with replacement) is taken into account for selecting instances in the random forest to build each tree. Each decision is therefore trained on distinct dataset based on different features.

5.2.7 Adaboost

AdaBoost is an ensemble of classifiers that uses simple classifiers, combines models of the same type, uses iterative model building, learns from the mistakes of previous models, and chooses the most popular model. This technique was very suitable to our dataset and it outperformed the results of the study conducted by the authors of the dataset [19].

5.2.8 PART

Described in the unsupervised section.

6 Performance Evaluation and Results

In this section we discuss the results and the findings of both supervised and unsupervised algorithms that were applied in this research project. We provide a set of charts and tables comparing the results of the different methods that we applied on the dataset. We first ran the methods on training sets, then evaluated them on test sets. Our training and testing datasets are split as follows: D3 (66% training, 34% testing), D4 (66% training, 34% testing), and D5 (50% training and 50% testing).

6.1 Unsupervised Learning

6.1.1 Association Rule Mining

Using Weka, we applied the Apriori algorithm and discovered a rule related to gender and the number of HCV cases found in the dataset such that if gender is “Female” then “No HCV is detected in the liver” with Confidence = 0.91, Support = 0.1 & Lift = 1.03. This rule is also supported by [25, 19]. In addition, the related papers suggest that men are found to have advanced stages of HCV compared to women due to life choices. “Several studies have shown an association of increased serum testosterone levels and advanced hepatic inflammatory activity in men” [25].

6.1.2 PART

We understand that PART is a supervised learning method, but we include its results in this section of unsupervised learning because we wanted to group the rules in one section including those generated by PART. Using either PART alone or AdaBoost with PART, we were able to create excellent set of rules for the HCV dataset with an accuracy greater than 95%. We list a few of these rules:

- if AST $\leq$ 52.3 and ALT $>$ 11.6 and Sex = Male and ALP $\geq$ 42.7 then Healthy Liver (100% accuracy)
- if BIL $>$ 9.6 then "HCV Liver" (100% accuracy)
- if CHE $\leq$ 13.71 and CREA $\leq$ 127 and AST $\leq$ 44.1 then Healthy Liver (274.89/1.42)
- if ALT $\leq$ 32.2 then HCV Liver (115.58/0.0)

Comparing our results with the work of Hoffman et al., we noticed that while that we outperformed their study by at least 10% (Hoffman's et al. study using PART with LOOV achieved only 85% accuracy).

6.2 Supervised Learning

6.2.1 Phase 1: Identifying Healthy vs. HCV Patient:

For this experiment, we have considered two variants. One is to drop the missing values using dataset D3 and another is to impute the missing values with corresponding feature mean using dataset D4. The performance of the classifiers applied to D3 were higher than the performance of the the supervised classifiers after dropping the missing values of D4. The results are showcased in Table 3 and Figure **??**. We also plotted an ROC Curve for HCV patients as shown in 4. From these 3 images, we can easily see that J48 and AdaBoost with PART outperform all the other supervised method. Both Decision Tree and the ensemble technique produce highest accuracy, recall precision for this HCV dataset. However, J48 had the least False Positives than AdaBoost.

Table 4 presents the performance metrics of the classifiers after mean imputation. Like the previous one, Adaboost for PART gives the best results. However, extent of the performance of mean imputation is comparatively low compare to the dropping missing values.

Comparing Table 3 with Table 4, we observe that the classifiers performed better on the dataset D3 (the one with no missing values) than on D4 (the one with mean imputed values).

Table 3: Performance Metrics for Supervised Classifiers After Dropping the Missing Value

Model	Accuracy	Recall	Precision
Logistic	0.96	0.97	0.97
Naive Bayes	0.89	0.91	0.91
KNN	0.91	0.91	0.91
Decision Tree (J48)	0.98	0.98	0.98
PART	0.95	0.95	0.96
Random Forest	0.95	0.78	0.97
AdaBoost (Logistic)	0.96	0.96	0.96
AdaBoost (Naive Bayes)	0.91	0.91	0.91
AdaBoost (KNN)	0.92	0.92	0.91
AdaBoost (J48)	0.96	0.97	0.97
AdaBoost (PART)	**0.98**	**0.98**	**0.98**

In a nutshell, J48 and Adaboost with PART perform best to identify both the HCV patients and the healthy subjects who qualify to become transplant donors for the HCV patient.

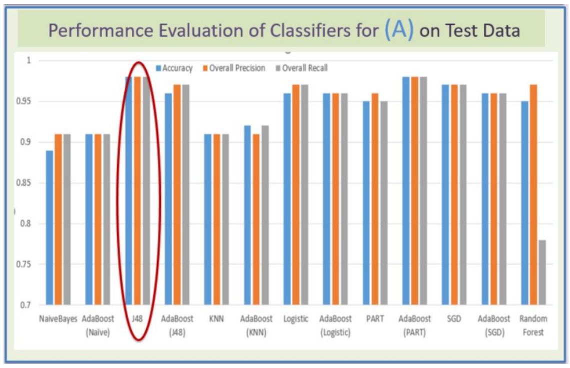


Figure 3: Performance Evaluation for Supervised Classifiers on D3

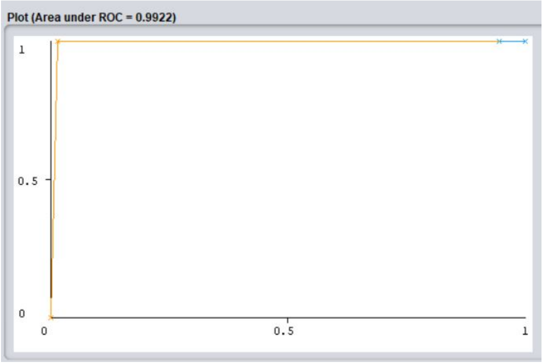


Figure 4: ROC Curve created by J48 on Dataset D3

Table 4: Performance Metrics for Supervised Classifiers After Imputing the Missing Values by Mean

Model	Accuracy	Recall	Precision
Logistic	0.96	0.96	0.96
Naive Bayes	0.87	0.87	0.87
KNN	0.9	0.9	0.9
Decision Tree (J48)	0.941	0.94	0.94
PART	0.948	0.94	0.94
Random Forest	0.94	0.77	0.96
AdaBoost (Logistic)	0.909	0.9	0.9
AdaBoost (Naive Bayes)	0.885	0.89	0.88
AdaBoost (KNN)	0.92	0.91	0.9
AdaBoost (J48)	0.96	0.96	0.96
AdaBoost (PART)	**0.97**	**0.97**	**0.97**

6.2.2 Phase 2: Identifying Cirrhosis Patients

It is commonly known that large number of instances in a dataset are a prerequisite for high accuracy in machine learning algorithms. Since our dataset is highly imbalanced, we tried to overcome this problem by creating two datasets D3 & D5. Dataset D5 includes instances of only HCV infected patients with the three known classes: Hepatitis, Fibrosis, & Cirrhosis. Since the size of D5 is too small (75 instances) for a machine learning algorithm, identifying cirrhosis patients is challenging due to the lack of data. Despite this fact, we executed the methods listed in Table 5 and generated the results based on accuracy, precision, and recall. It is evident that AdaBoost using either J48 or PART outperforms the other classifier, also shown in Figure 5.

7 Conclusion and Future Work

In this work, we set a goal to identify HCV infected patients based on features obtained from various blood test results. We intended to identify different stages of liver disease on identified patients. From the available features we had, AST, ALP, BIL and ALT proved to be the most informative ones.

We have been able to identify healthy and HCV infected livers with great accuracy. We have also been able to identify the different stages of the infected liver disease, in order to determine the patients that qualify for liver transplant, and to identify potential donors. From all the methods that we applied in this work, we observe that J48 works best for identifying healthy donors. We also observe that AdaBoost with PART performed best in identifying Cirrhosis for potential recipients of liver transplant.

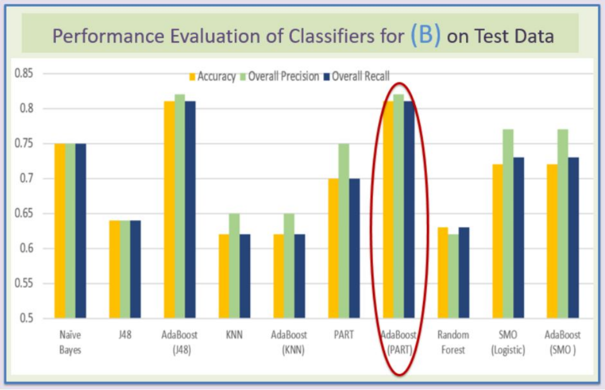


Figure 5: Performance Evaluation Chart Using Supervised Classifiers on Dataset D5

Table 5: Performance Metrics for Identifying Cirrhosis Patient

Model	**Accuracy**	**Recall**	**Precision**
Logistic	0.72	0.73	0.77
Naive Bayes	0.75	0.75	0.75
KNN	0.62	0.62	0.65
Decision Tree (J48)	0.64	0.64	0.64
PART	0.7	0.7	0.75
Random Forest	0.63	0.63	0.62
AdaBoost (SMO)	0.72	0.73	0.77
AdaBoost (KNN)	0.62	0.62	0.65
AdaBoost (J48)	**0.81**	**0.81**	**0.82**
AdaBoost (PART)	**0.81**	**0.81**	**0.82**

To complete this work we have to overcome some challenges. Among which, working on a small dataset was the most difficult one. Even in this small dataset, there are a lot of missing values. Due to time limit, we could not treat the missing value problem fully (we performed two out of three methods for missing values). In the future, we plan to predict the missing values by creating a classifier for each column and replace each missing one by the predicted values. We expect this to generate better results. In addition, the classes (stages of HCV and no HCV) are highly imbalanced. The lack of a bigger and balanced dataset might have affected the performance of machine learning models. In future, we intend to improve the accuracy by using data augmentation techniques for highly imbalanced classes.

Acknowledgments

We would like to acknowledge Dr. James Foulds for his guidance especially during this COVID-19 time. In addition, we are gratefully thankful to the creators of the dataset Drs. Lichtinghagen et al. who responded to our emails and clarified various aspects of the dataset.

References

[1] Léon Bottou. "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[2] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] CDC. *Division of Viral Hepatitis.* 2020. URL: https://www.cdc.gov/hepatitis/index.htm.

[4] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.

[5] Mayo Clinic. *Liver Transplant.* 2016. URL: https://www.mayoclinic.org/tests-procedures/liver-transplant/about/pac-20384842.

[6] Lab Corp. *SI Conversion Unit Table.* 2020. URL: https://www.labcorp.com/resource/si-unit-conversion-table.

[7] Quest Diagnostics. *Quest Diagnostics Test Directory.* 2000. URL: https://testdirectory.questdiagnostics.com/test/home.

[8] O Doyle, N Leavitt, and J Rigg. *Finding undiagnosed patients with hepatitis C infection: an application of artificial intelligence to patient claims data.* 2020. URL: https://doi.org/10.1038/s41598-020-67013-6.

[9] Eibe Frank and Ian H Witten. "Generating accurate rule sets without global optimization". In: (1998).

[10] Mark Hall et al. "The WEKA data mining software: an update". In: *ACM SIGKDD explorations newsletter* 11.1 (2009), pp. 10–18.

[11] Jun Han and Claudio Moraga. "The influence of the sigmoid function parameters on the speed of backpropagation learning". In: *International Workshop on Artificial Neural Networks.* Springer. 1995, pp. 195–201.

[12] *Hepatitis C.* https://www.cdc.gov/hepatitis/hcv/index.htm. Accessed: 2020-12-10.

[13] Tapas Kanungo et al. "An Efficient K-Means Clustering Algorithm: Analysis and Implementation". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 24.7 (July 2002), pp. 881–892. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2002.1017616. URL: https://doi.org/10.1109/TPAMI.2002.1017616.

[14] M Khan et al. *A machine-learning algorithm to identify hepatitis C in health insurance claims data.* 2019. URL: https://journals.uic.edu/ojs/index.php/ojphi/article/view/9685.

[15] David G Kleinbaum et al. *Logistic regression.* Springer, 2002.

[16] S Kotsiantis and D Kanellopoulos. *Data Preprocessing for Supervised Leaning.* 2007. URL: https://publications.waset.org/14136/pdf.

[17] Sotiris Kotsiantis and Dimitris Kanellopoulos. "Association rules mining: A recent overview". In: *GESTS International Transactions on Computer Science and Engineering* 32.1 (2006), pp. 71–82.

[18] Ralf Lichtinghagen and Frank Klawonn. *UCI Machine Learning Repository, HCV data Data Set.* 2020. URL: http://archive.ics.uci.edu/ml/datasets/HCV+data.

[19] Ralf Lichtinghagen and Frank Klawonn. *Using machine learning techniques to generate laboratory diagnostic pathways—a case study.* 2018. URL: http://jlpm.amegroups.com/article/view/4401/5425.

[20] Mohammed Al-Maolegi and Bassam Arkok. "An improved apriori algorithm for association rules". In: *arXiv preprint arXiv:1403.3948* (2014).

[21] Todd K Moon. "The expectation-maximization algorithm". In: *IEEE Signal processing magazine* 13.6 (1996), pp. 47–60.

[22] NIH. *Compound.* 2020. URL: https://pubchem.ncbi.nlm.nih.gov/.

[23] NIH. *The Natural History of Hepatitis C Virus (HCV) Infection.* 2006. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1415841/.

[24] William S Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[25] Hepatitis C Online. *Core Concepts - Evaluation, Staging, and Monitoring of Chronic Hepatitis C.* 2020. URL: https://www.hepatitisc.uw.edu/go/evaluation-staging-monitoring/evaluation-prognosis-cirrhosis/core-concept/all.

[26] World Health Organization. *Stages of Liver Disease.* 2016. URL: https://www.slideshare.net/Moheer07/hepatitis-c-and-its-treatment.

[27] L. E. Peterson. "K-nearest neighbor". In: *Scholarpedia* 4.2 (2009). revision #137311, p. 1883. DOI: 10.4249/scholarpedia.1883.

[28] J. R. Quinlan. "Induction of Decision Trees". In: *Mach. Learn.* 1.1 (Mar. 1986), pp. 81–106. ISSN: 0885-6125. DOI: 10.1023/A:1022643204877. URL: https://doi.org/10.1023/A:1022643204877.

[29] Irina Rish et al. "An empirical study of the naive Bayes classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence.* Vol. 3. 22. 2001, pp. 41–46.

[30] Steven L Salzberg. *C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.* 1994.

[31] Robert E Schapire. "Explaining adaboost". In: *Empirical inference.* Springer, 2013, pp. 37–52.

[32] US Department of Veteran Affairs. *The Liver is a filter.* 2019. URL: https://www.hepatitis.va.gov/basics/liver-as-filter.asp.

[33] Ian H Witten et al. *Data Mining, Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2000.