A STUDY TO PREDICT THE STAGES OF LIVER FIBROSIS CAUSED BY HEPATITIS C VIRUS USING DATA MINING TECHNIQUES

BY

MD HASIN ISHRAQ
ID: 191-25-728

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By

Md. Sazzadur Ahamed
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

DECEMBER 2020

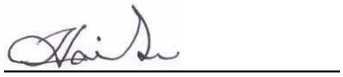
APPROVAL

This Thesis titled "**A Study to Predict the Stages of Liver Fibrosis Caused by Hepatitis C Virus Using Data Mining Techniques**", submitted by **Md Hasin Ishraq; ID: 191-25728** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 22 December 2020.

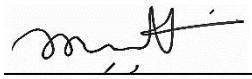
BOARD OF EXAMINERS

Dr. Syed Akhter Hossain Chairman Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Ismail Jabiullah Internal Examiner Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Sheak Rashed Haider Noori Internal Examiner Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology Daffodil
International University



Dr. Mohammad Shorif Uddin External Examiner Professor
Department of Computer Science and Engineering
Jahangirnagar University

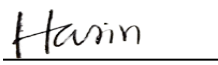
DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Md. Sazzadur Ahamed, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Sazzadur Ahamed
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md. Hasin Ishraq
ID: 191-25-728
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty ALLAH for His divine blessing makes us possible to complete the master's thesis successfully.

We really grateful and wish our profound indebtedness to **Md. Sazzadur Ahamed**, Senior Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Data Mining and Machine learning" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

We would like to express our heartiest gratitude to **Prof. Dr. Syed Akhter Hossain** Head of the Department of CSE, for his kind help to finish our thesis and also to other faculty member and the staff of CSE department of Daffodil International University.

We are grateful to **Dr. ASM Shahid Ullah** Vice Principle of PAHMC, BGen(retd) **Dr. AFM Wasy;** CMH Dhaka, **Dr. Shamima Sultana** Professor Pharmacology & **Dr. Shyamal Kumar Paul** Professor Microbiology of MMC.

Special thanks to **Lamia Sultana**; medical student who gave us a lot of support with medical books.

ABSTRACT

Hepatitis is basically an inflammation of liver. There many forms of hepatitis virus. Among them, hepatitis B as well as hepatitis C viruses lead to liver cancer. Hepatitis C virus (HCV) is the notorious toxicant of liver. HCV harms the liver upon time. There reside mainly four stages of liver fibrosis by which the liver gets damaged. Those stages are portal fibrosis, portal fibrosis with few septa, numerous septa without cirrhosis, cirrhosis. At the first stage, the liver gets inflamed. At the time of fibrosis, defected tissue starts replacing cured tissue in the inflamed liver. When fibrosis of liver begins various scarring starts building up and make it tough for the liver to work thoroughly. HCV includes some symptoms like fever, nausea or vomiting, headache, Diarrhea, fatigue, jaundice, epigastric pian. Now a days, the number of HCV infectious individuals is growing globally which has become a matter of concern. In order to save lives, researchers have over the years worked to discover an alternative diagnostic means for hepatitis disease using computing intelligence. An early diagnose as well as prediction of liver disease like hepatitis is quite beneficial. Computing intelligence is playing a significant role for the realm of healthcare now a days. Our objective is to compute the ratio of stages of liver fibrosis and the symptoms of hepatitis C virus. In this paper, with the help of data mining the stages of fibrosis has been classified. Three significant algorithms Artificial Neural Network (ANN) Naive Bayes (NB) as well as J48 were used in this paper for the stage prediction of fibrosis of liver. All of those three efficient algorithms performed quite well but among them, J48 decision tree algorithm gave a paramount performance with an accuracy of 96.03%. By Utilizing J48 algorithm it is possible to classify the different stages of liver fibrosis sophisticatedly as well as to develop an expert system in future work. So J48 is considered the most convenient algorithm for classification of the stages of liver fibrosis caused by hepatitis C virus as well as their early diagnosis. This datamining process will also help to avoid the hazard of invasive procedure for diagnosis of liver fibrosis.

TABLE OF CONTENTS CONTENTS PAGES

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Introduction

If we talk about crucial organs in the human anatomy, then liver will be one of them, its weight is around 1500g (1.5 kg). Liver is reddish brown in color and it is separated into four lobes of irregular shape and size. Various conditions can strafe liver tasks as well as their malady can range from basic drug reactions to liver disease. The liver works a variety of important functions relating to metabolism, digestion, immunity as well as the conservation of nutrients within the body. These mentioned functions the liver act as an essential organ, without which body tissues will rapidly die from a lack of energy and nutrients [1]. Hepatitis C virus which is also called HCV is responsible for liver disorder which has become a global problem. HCV may lead to serious morbidity as well as mortality in the future. Hepatitis C ultimately results fibrosis of liver. Which has five stages; which can be denoted from F0 to F4. F0 stage is called no fibrosis. F1 stage is called portal fibrosis. F2 stage is called portal fibrosis with few septa. F3 stage is called numerous septa without cirrhosis. F4 stage is called cirrhosis. Till now vaccine is unavailable for HCV, the main prevention exertion should concentrate on cleaner blood supplies, healthy injecting procedures, special attention should be given during surgical procedure [2][3]. Data mining is originally a method of information exploration from a broad database and consists of many various algorithms and explained computational techniques. Both of these data mining strategies are applied to classify efficient tidings from a massive compilation of various types of data. Mined knowledge is interpreted like a representation about the semantic formation of the data collection and the sampling perhaps applied to forecast and identify novel data [4].

In this paper, by the help of a data mining algorithm the stages of liver fibrosis caused by hepatitis C virus have been predicted. The objective of this paper is to prospect a pathway for the enumeration of several stages of liver fibrosis resulted from hepatitis C virus based on symptoms and laboratory tests.

1.2 Motivation

Hepatitis C is a worldwide medical condition as the World Health Organization, announced 3-4 million individuals are recently contaminated by hepatitis C infection each year as well as 130-170 million individuals are tainted. More than 350000 individuals kick the bucket every year from hepatitis C-narrated liver disorders [5]. Over 60% HCV-tainted individuals in Bangladesh are somewhere in the range of 30 and 50 years old, and HCV is the subsequent driving reason for Liver cirrhosis in Bangladesh, representing 30% instances of liver cirrhosis and 17% instances of HCC (Hepatocellular Carcinoma) [6]. When people are affected by hepatis C virus, it can cause cirrhosis. As HCV advances, signs like skin issues, blood issues, and weight reduction may show up. Risky results like serious liver harm, liver malignant growth, and liver disappointment can happen. It is very important to detect or predict the stages. It is possible to make predictions of the stages of fibrosis caused by hepatitis C virus; by applying various datamining algorithms as well as data excavation. Socio economic condition of most of the people of our country is not up to the mark. So, they can't avail or purchase proper medical benefit. For diagnosis of liver fibrosis, it's required to perform Needle puncher or USG FNAC (Ultra sound guided Fine Needle Aspiration Cytology) test. Not all medical personnel, doctor can perform these tests. Only the specialist doctor can conduct these tests. These tests are costly. These tests are invasive in nature, which will cause pain, injury, infection. Doctor may be contaminated with HCV from patient during performing these procedures. The patient may also be contaminated with other infectious diseases by the surgical instruments. It may also damage associative organs like- gallbladder, bile duct etc. For avoiding these untoward effects of diagnostic procedure and high cost has motivated us to perform this work.

1.3 Research Questions

In this research we will be trying to solve the following questions:

1. Which algorithm would be best among Naïve Bayes, J48 & ANN for predicting the stages of liver fibrosis caused by hepatitis C virus?
2. What are the ratio of each symptom with the stages of liver fibrosis caused by hepatitis c virus?

1.4 Expected Output

Expectation isn't generally precise and some of the time it will give erroneous outcomes in various circumstances by investigating the information. From this work, we expect our investigation will assist to predict the stages of liver fibrosis caused by hepatitis C virus. After training our dataset by different algorithms we hope to have good accuracy of our testing dataset.

1.5 Report Layout

In the first chapter of the thesis, we have discussed the overview of the thesis, our motivation of the thesis. We have also discussed the expected outcome that we will get. The second chapter we have discussed our background study on hepatitis C virus and the stages of liver fibrosis which contains introduction, related work, research summary, scope of the problem and challenges. The third chapter is about the research methodology we have applied. Here we have discussed about the dataset and it's preprocessing and processing, our study work flow, used datamining techniques and research instruments. The fourth chapter is about the experimental results of our thesis. In the fifth chapter we have discussed the summary, future scope of the study, recommendations and conclusion.

CHAPTER 2 BACKGROUND

2.1 Introduction

Liver is a wedge-shaped limb and it's the largest gland of the body. It Occupied much of the right upper part of the abdominal. It is 2.5% of adult body weight. The liver does not contain only one feature. Liver perform so many tasks, some of them are- storing glycogen, synthesizing plasma protein, lipids, detoxifying blood. Secretion of bile from bile duct is also done by liver. When a person gets infected by the hepatitis virus (HCV), the virus may invade his liver and it causes inflation and redness in the liver. When the liver gets contaminated by an infection or gets affected by toxins the liver gets too weakened and it gets unable to function and to keep a human alive. Hepatotropic virus-induced liver disease puts a significant burden on health care services [1]

Hepatitis is referred to as inflammation (itis) of the liver (hepar). There are numerous reasons for hepatitis including A, B, C, D as well as E viral infections. Hepatitis A and

hepatitis E transmitted through contaminated food or water and hepatitis B, C and D are transmitted through blood and body fluids. HCV discovered in 1989. This is estimated that about 110 million persons are infected by HCV globally. Hepatitis C virus (HCV) is basically RNA virus. Fever, Nausea or vomiting, jaundice, Headache, Diarrhea, Epigastric pain and fatigue are common symptom of HCV. These symptoms may vary from person to person, either one or more of the symptoms can be exhibit in a patient body even none of the symptoms can't be seen. Infection with HCV is exceedingly asymptomatic with no external symptoms, approximately 80% of the affected people will not exhibit the symptoms initially. Most severe infections lead to chronic, followed by liver ailments like cirrhosis and hepatocellular carcinoma. In a patient body if there is presence of antibody against HCV then we can say that person is infected. HCV frequently transmitted by exposure of infected blood or organ, contaminated instruments like- needle and syringe, drug which needed to be injected, infant; if mother is affected. Tests like- WBC, RBC, HGB, Plat, AST, ALT, RNA Base, RNA EOT, RNA EF are used to diagnose liver disease. And for the confirmation of diagnosis needle puncher or Ultra sound guided FNAC is required. Early diagnosis can help to prevent from liver cirrhosis and liver cancer. Its recommended that the following types of people may be at risk of infection, such as- if a patient receive organ or blood, injected drug user, kidney dialysis patient, health care workers, HIV patient, liver disease patient, infant; whose mother is HCV infected, people who has gone under invasive operation, people who has HCV infected sexual partner, who has tattoos and piercing [7],[8].

Hepatitis C virus generally affects the liver. Because, when HCV enter in the blood, it tries to find a suitable place for itself. So, the viruses search for the preferable place in the whole body. When these viruses enter the liver, they find out that the liver cell is desirable place they are looking for. After finding the suitable place the hepatitis C virus starts to attack the liver cells and thus the life cycle of HCV begins.

Life cycle of HCV in a liver cell is described below.

(1) Hepatitis C Virus: HCV is mainly formed by RNA and protein. There is also presence of glycoprotein and lipid membrane.

(2) Attachment: At this stage, liver start to attach with the cell.

(3) Penetration and Entry: Cell membrane is formed with phospholipids bi layer. There is also glycolipid, glycoprotein. As we know HCV has glycoprotein and lipid membrane, so as a result HCV can easily do penetrate the cell membrane in disguise. So, at this point cell cannot detect the threat.

(4) Fusion and virus RNA release: Here, HCV fuse with the cell and release the RNA of it.

(5) Production of protein: Normally cell generates protein with the help of RNA. When HCV fuse with the cell it takes over the process and its RNA start the production of protein.

(6) Protein processing: In this step, protein is being processed in the cell.

(7) Replication: By the polymerase enzyme so many copies of RNA are made.

(8) Viral assembly: Those replicated RNA starts to collect its essential elements from the cell to generate new hepatitis C Virus. Protein shell starts to build around the RNA of HCV.

(9) Secretion: In this step the HCV is preparing itself to get out of the cell.

(10) Release: The new viruses of hepatitis C are released from the infected cell.[9]

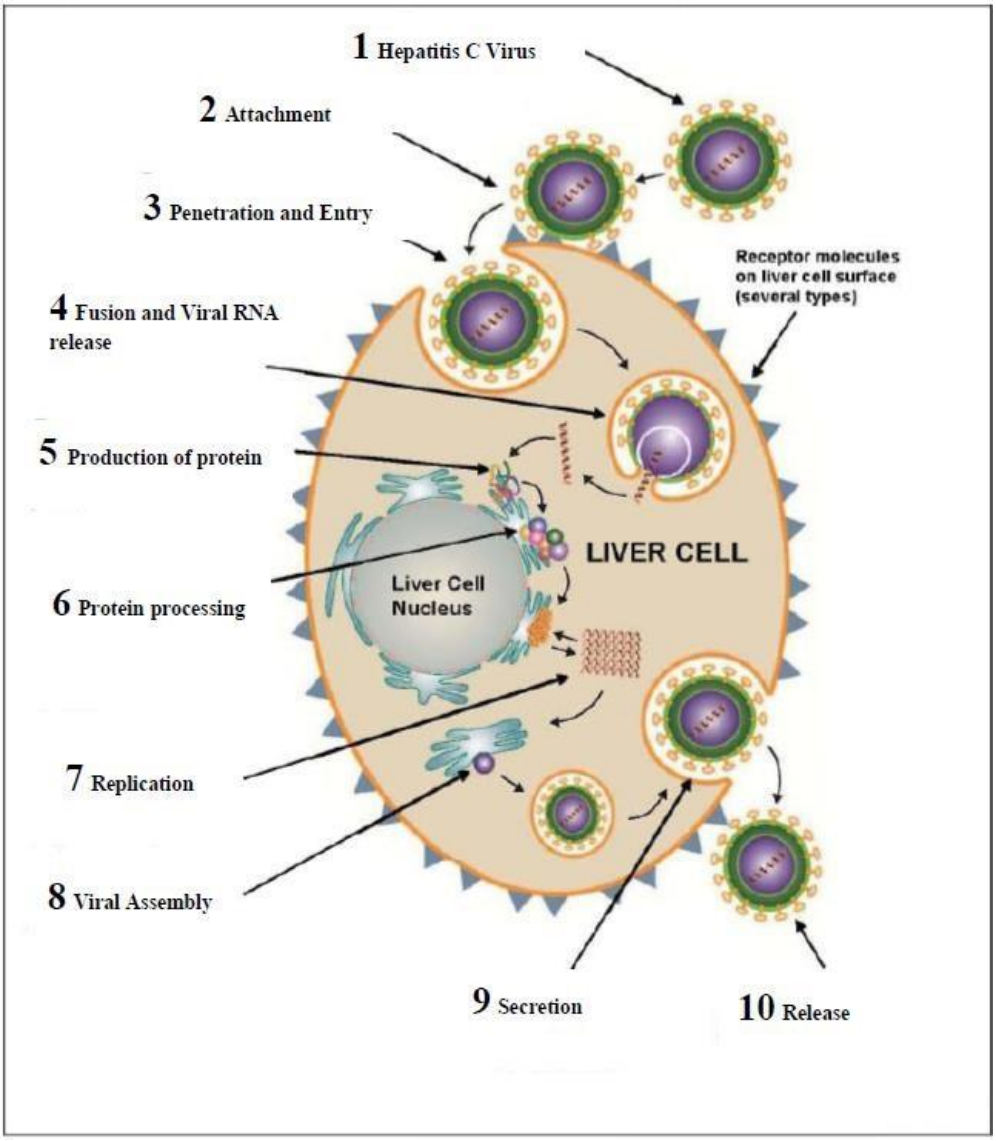


Figure 2.1: Life cycle of HCV

Hepatocyte is the cell of liver. When liver get to know that it's being affected by viral infection (HCV) as an immune response hepatocyte release Cytokine (chemical). TNF (Tumor Necrosis Factor) is a type of cytokine. TNF stimulates the stellate cell. As necessary, stellate cell can convert into necessary elements. In this case stellate cell converts into fibroblast. Fibroblast release fiber, which is called collagen fiber. So as a result, it causes fibrosis to repair the damage. Fibrosis heals the wound and scar tissue is formed. As a result, the wound is healed but the scar tissue remains there. Ultimately large amount of abnormal scar tissue in the liver is formed for fibrosis and at last severe scaring

can result cirrhosis. Fibrosis is called an accumulation of scar tissue in the liver. It can obstruct the flow of blood to liver cells and hamper liver's function. The liver cells pass over time and the liver does not work normally anymore. The body can no longer retain the weakening liver as further damage continues. At the very beginning, the body works its best for offsetting for bad liver function. But in respect of time, the liver turns out so faulty that it doesn't function appropriately and It gets unable to perform its signification actions for the body. People who are affected by cirrhosis may have complications like bleeding, easy bruising, exhaustion, annoyance, itchiness. In liver there is portal triad, central vein, septa (septum is singular form of septa) etc. Portal triad consists of Hepatic portal vein, portal hepatic artery and common hepatic duct (bile duct). In liver there are many lobules, these are called hepatic lobules. Lobule of a liver consists of so many hepatocytes [8],[10].

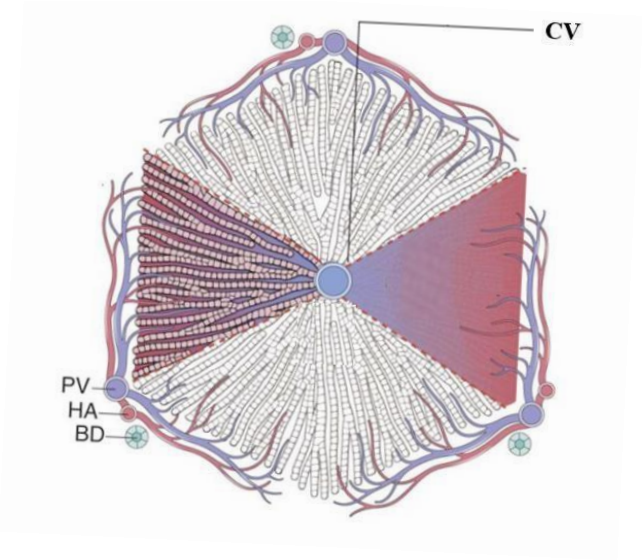


Figure 2.2: Model of hepatic lobule

Stage 0 (No Fibrosis): It's the normal stage of a liver. It's not affected by hepatitis C virus; till now there is no scar tissue formed for fibrosis.

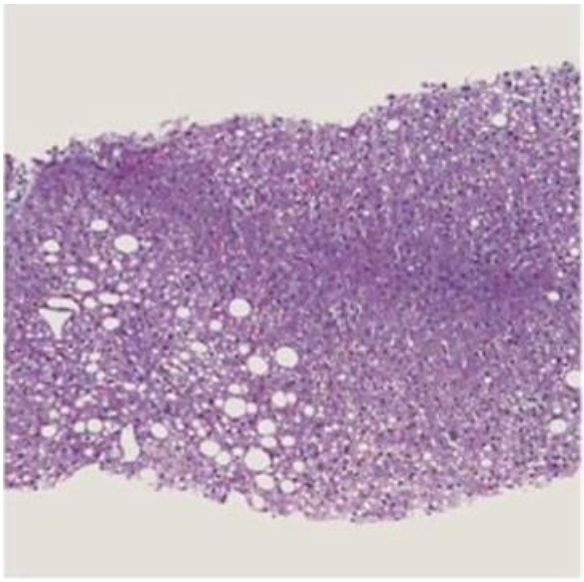


Figure 2.3: Microscopic view of liver tissue

Stage 1(Portal Fibrosis): When HCV attacks, first it will start to damage the portal regions of the lobule. So as a result, it causes fibrosis to repair the damage of the portal triads.

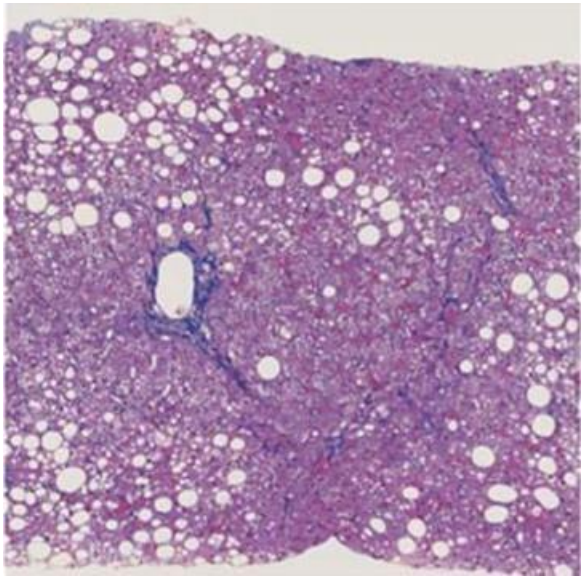


Figure 2.4: Microscopic view of liver fibrosis stage 1

Stage 2 (portal fibrosis with few septa): Septum mean partition that separate one thing from another. In a lobule there are so many cells. These cells are separated from each other by septa. Septa are like a membrane. In the lobule the cells are organized as spoke. This cell spokes are also separated from one another by septa. When septa are attacked by hepatitis

C virus; HCV viruses start to damage the septa. To protect the septa fibrosis occurs. When low amounts of septa get affected by HCV for which fibrosis occurs; this stage is called few septa.

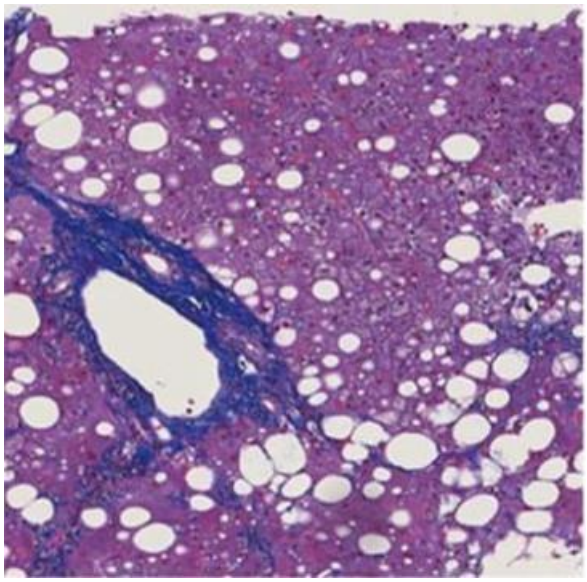


Figure 2.5: Microscopic view of liver fibrosis stage 2

Stage 3(Numerous Septa Without Cirrhosis): When numerous numbers of septa are affected with fibrosis it's called numerous septa.

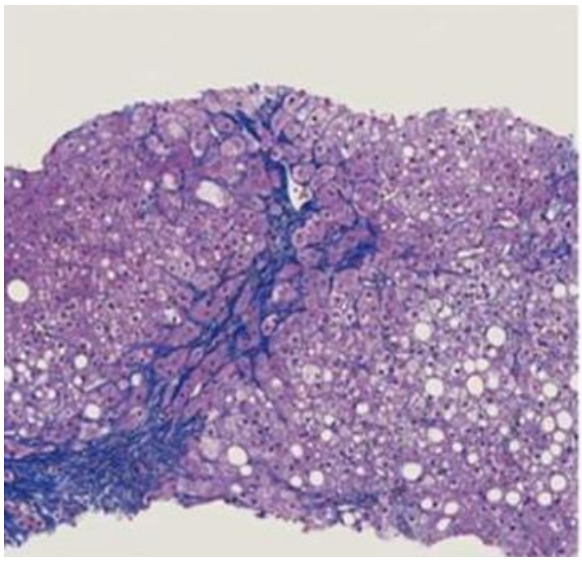


Figure 2.6: Microscopic view of liver fibrosis stage 3

Stage 4(Liver Cirrhosis or Total Septa): Like the above stages, when almost whole liver got affected with the fibrosis; this stage is called Liver Cirrhosis.

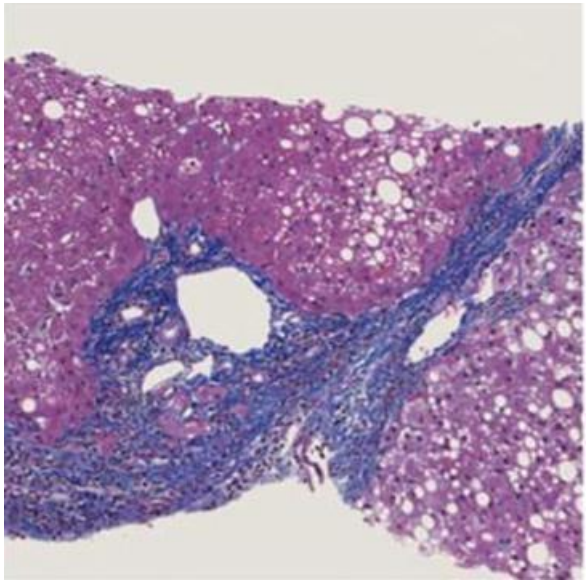


Figure 2.7: Microscopic view of liver fibrosis stage 4

All these processes happen in other lobules also [10],[11].

If cirrhosis is not detected early; it can cause so many complications; like- vein that supply blood to liver there will be high blood pressure; which is called portal hypertension, ascites; for which abdomen and leg will swell, rupture of the vein, infection, interruption to processes nutrition, blood toxin can't n be removed by the damaged liver as a result toxin start to build in the brain, lose the strength of the bone. Ultimate complication is cirrhosis causes HCC (Hepatocellular carcinoma) for which death is the ultimate result [12].

The involvement of viruses in the liver is the major cause in hepatitis disease. Hepatitis is a condition with high death rates worldwide. Hepatitis can lead to cirrhosis, serious scarring as well as raise the chance of liver cancer if precise steps are not taken in due time. Early detection can cure the disease through appropriate diagnosis as well as accurate provision. The two important and significant things for the diagnosis of any disease are (i) selecting the appropriate diagnostic parameters and (ii) properly analyzing the data with experienced and sophisticated sagacity. Machine Learning (ML) is the exact arsenal that can allow a

device to use various algorithms to acquire knowledge by itself by identifying attribute for the conferred data. This allows automated diagnosis of any diseases [17].

A significant percentage of infected people by hepatitis are entirely ignorant of their conditions as well as protocols for prevention. The deficiency of sufficient treatment services, low economic status, inexperienced health care personnel, and lack of knowledge are the main reason behind hepatitis, nearly incurable as well as too costly that such an investment can't be afforded by a poor citizen. As there are no vaccines available yet and acquainted treatment for hepatitis c virus has yet to be discovered. Moreover, owing to the cost of treating liver disease, hepatitis is a weighty compulsion on the health care system. Many affected patients can be rescued by early prediction and accurate disease detection. In this paper, several algorithms are used for the prediction of the stages of liver fibrosis caused by hepatitis C virus (HCV).

2.2 Related Work

Radwan et al. developed a system that represents several types of classification methods in data excavation for forecasting the infection with HCV from the HCV data set. They had used three algorithms namely Decision Tree, Naïve Bayes and Neural Network. In the 1800 record dataset size the accuracy of the Decision Tree and Naïve Bayes [14].

Nandipati et al. proposed execution examinations to predict the hepatitis C virus from there study. For the research they have use ML for projection. After completing their study, they have found that; the precision by KNN (51.06%) and for random forest they have achieved (54.56%) precision. [15].

Yada N. et al utilized decision tree to perform their study on liver fibrosis which is caused by HCV. For stage 1 of liver fibrosis, they have found precision (94.4%); for stage 2 (54.1%), for stage 3(38.7%); for stage 4(81.3%). F1, F2, F3, and F4 of, 54.1, 38.7, and 81.3%. [16].

Heba Ayeldeen et al proposed a model based on decision tree. They tried to perform their study to predict the stages of liver fibrosis. After completing their research, they have found precision 93.7% by using decision tree algorithm [17].

Brian Keltch et al have worked with both HCV and HBV. For performing their study, they have utilized four techniques, among they found best result using decision tree. From the study they made a model using the algorithm and from that they have also made a prototype [18].

Nahla H. Barakat et al have utilized random forest in their work. They have worked with children; who has liver fibrosis. They build a system to predict the stages of liver fibrosis. They have achieved precision of 87.5% from their study [19].

2.3 Research Summary

The aim of this research is to predict the stages of liver fibrosis caused by HCV. Hepatitis C virus is a significant reason for liver disease. At the beginning of the work dataset will be collected then we will analyze the data. We will use different types of algorithms in our data. We will see how much our models can efficiently predict the stages of liver fibrosis resulted from hepatitis c virus. There three algorithms will be used here; namely J48, Artificial Neural Network (ANN) and Naïve Bayes. We will try to utilize these algorithms hope fully with good accuracy in the work to predict the stages of fibrosis of liver which is caused by HCV; to avoid the invasive procedures. We will also try to observe the ratio (percentage) of each symptom with the stages of liver fibrosis resulted from HCV.

2.4 Scope of the Problem

Liver is a crucial organ in the human anatomy and the largest gland in the body. Liver may be contaminated by hepatitis infections for which fibrosis occur in the liver. Fibrosis can be resulted from numerous kinds of infections; among them we will attempt to play out our examination on hepatitis C infection. Since after hepatitis B infection; hepatitis C infection is the new emerging hepatitis infection in Bangladesh. So now it's a greater worry for the individuals of Bangladesh. In our exploration we will be attempting to predict the stages of liver fibrosis caused by hepatitis C virus. On the off chance that we can prevail in our exploration, later on it will likewise help different analysts for their further examinations to secure more successful outcomes.

2.5 Challenges

The main challenge of our study is having the proper prediction and accuracy using the algorithms. We have to customize some of the parameters of our algorithms on the WEKA. If we start utilizing without processing the data, the accuracy may decrease. We can use different tools, algorithms, library functions to process our data set and we needed to do proper distribution of training and testing data so that we can predict in a fruitful way. For our study we needed to collect authentic information; initially it gave a lot of trouble. For acquiring authentic information, we have searched for scarce medical books. We have to take appointment from senior medical personnel for gathering basic knowledge regarding our topic; like- HCV, invasive procedures, staging of liver cirrhosis; and also, to verify our dataset.

During our work Covid-19 situation created great troubles; PC couldn't be repaired for so many days for lockdown which wasted our time. And also; another extra burden was to cope up with new software at the last moment. Sometimes working in online takes much more time rather than doing the work in offline.

We are learner, we didn't have previous experience; there is no work exists without challenges, challenge comes from problem so we should have to take challenges for solving the problem.

CHAPTER 3 DATA COLLECTION AND ANALYSIS

3.1 Introduction

Research methodology is the way to collect and gather information and data for the purpose of making any decision. For prediction-based research; classification algorithms of machine learning (ML) are used to predict the outcome. We have used supervised learning technique and used J48, Naïve bayes and ANN algorithms for prediction of staging of liver fibrosis caused by hepatitis C virus in our study.

3.2 Data set

The dataset contains 29 attributes and among them one is class attribute. It contains 1385 instances. The attributes are shown in Table 3.1.

Table 3.1: Features List

Serial	Features Name	Sub Category	Data Distributions
1.	Age	Minimum: 32	Mean: 46.144
		Maximum: 61	STDev: 8.834
2.	Gender	Male	635
		Female	750
3.	BMI(Body Mass Index)	Minimum: 22	Mean: 28.654
		Maximum: 35	STDev: 4.123
4.	Fever	Present	557
		Absent	828
5.	Nausea/Vomiting	Present	559
		Absent	826

Serial	Features Name	Sub Category	Data Distributions
6.	Headache	Present	568
		Absent	817
7.	Diarrhea	Present	770
		Absent	615
8.	Fatigue	Present	714
		Absent	671

9.	Jaundice	Present	682
		Absent	703
10.	Epigastric pain	Present	570
		Absent	815
11.	WBC(White Blood Cells)	Minimum: 3029	Mean: 7991.928
		Maximum: 12093	STDev: 2954.2
12.	RBC(Red Blood Cells)	Minimum: 3820864	Mean: 4423309.109
		Maximum: 5012941	STDev: 355677.781
13.	HGB (Hemoglobin)	Minimum: 10	Mean: 12.825
		Maximum: 15	STDev: 1.672
14.	Plat(Platelet)	Minimum: 93731	Mean: 160468.66
		Maximum: 226464	STDev: 37611.794
15.		Minimum: 39	Mean: 84.59

Serial	**Features Name**	**Sub Category**	**Data Distributions**
	AST1(Aspartate Transaminase 1 Week)	Maximum: 128	STDev: 26.982
16.	ALT 1(Alanine Transaminase 1 Week)	Minimum: 39	Mean: 78.952
		Maximum: 128	STDev: 28.621
17.	ALT4(Alanine Transaminase	Minimum: 39	Mean: 80.979

	4 Week)	Maximum: 128	STDev: 24.777
18.	ALT 12(Alanine Transaminase 12 Week)	Minimum: 39	Mean: 77.916
		Maximum: 128	STDev: 24.413
19.	ALT 24(Alanine Transaminase 24 Week)	Minimum: 39	Mean: 75.482
		Maximum: 128	STDev: 29.289
20.	ALT 36(Alanine Transaminase 36 Week)	Minimum: 39	Mean: 79.617
		Maximum: 128	STDev: 26.922
21.	ALT 48(Alanine Transaminase 48 Week)	Minimum: 39	Mean: 82.799
		Maximum: 128	STDev: 28.961
22.	ALT 52(Alanine Transaminase 52 Week)	Minimum: 39	Mean: 33.238
		Maximum: 128	STDev: 6.891
23.	RNA Base	Minimum: 385	Mean: 564587.763
		Maximum: 1198310	STDev: 360378.506
24.	RNA 4	Minimum: 190	Mean: 634054.777
		Maximum: 1199513	STDev: 353711.393
Serial	**Features Name**	**Sub Category**	**Data Distributions**
25.	RNA 12	Minimum: 5	Mean: 276005.997
		Maximum: 801981	STDev: 265587.295
26.	RNA EOT	Minimum: 5	Mean: 239135.883

		Maximum: 808142	STDev: 252860.866
27.	RNA EF (Elongation Factor)	Minimum: 5	Mean: 285645.077
		Maximum: 807660	STDev: 279930.694
28.	Grading	Minimum: 3	Mean: 9.096
		Maximum: 16	STDev: 3.653
29.	Staging of liver cirrhosis caused by HCV (Class)	Portal Fibrosis	336
		Portal Fibrosis with Few Septa	332
		Numerous Septa without Cirrhosis	355
		Cirrhosis	362
STDev = Standard Deviation			

3.2.1 Data Collection Procedure

The source of the dataset we used for our work; has been prepared at Ain Shams University of Egypt. The dataset has been collected from UCI Machine Learning Repository.

3.2.2 Data Preprocessing

The raw data set contains all numerical value. Some of the attributes like Gender, Fever, Nausea/Vomiting, Headache, Diarrhea, Fatigue, Jaundice and Epigastria pain were converted to nominal value and rest attributes remain in numerical format. The CSV (Comma Separated Value) was converted to ARFF (Attribute-Relation File Format) file. This ARFF file contains all the attributes and instances of the dataset. This file format was designed at Department of Computer Science of The University of Waikato.

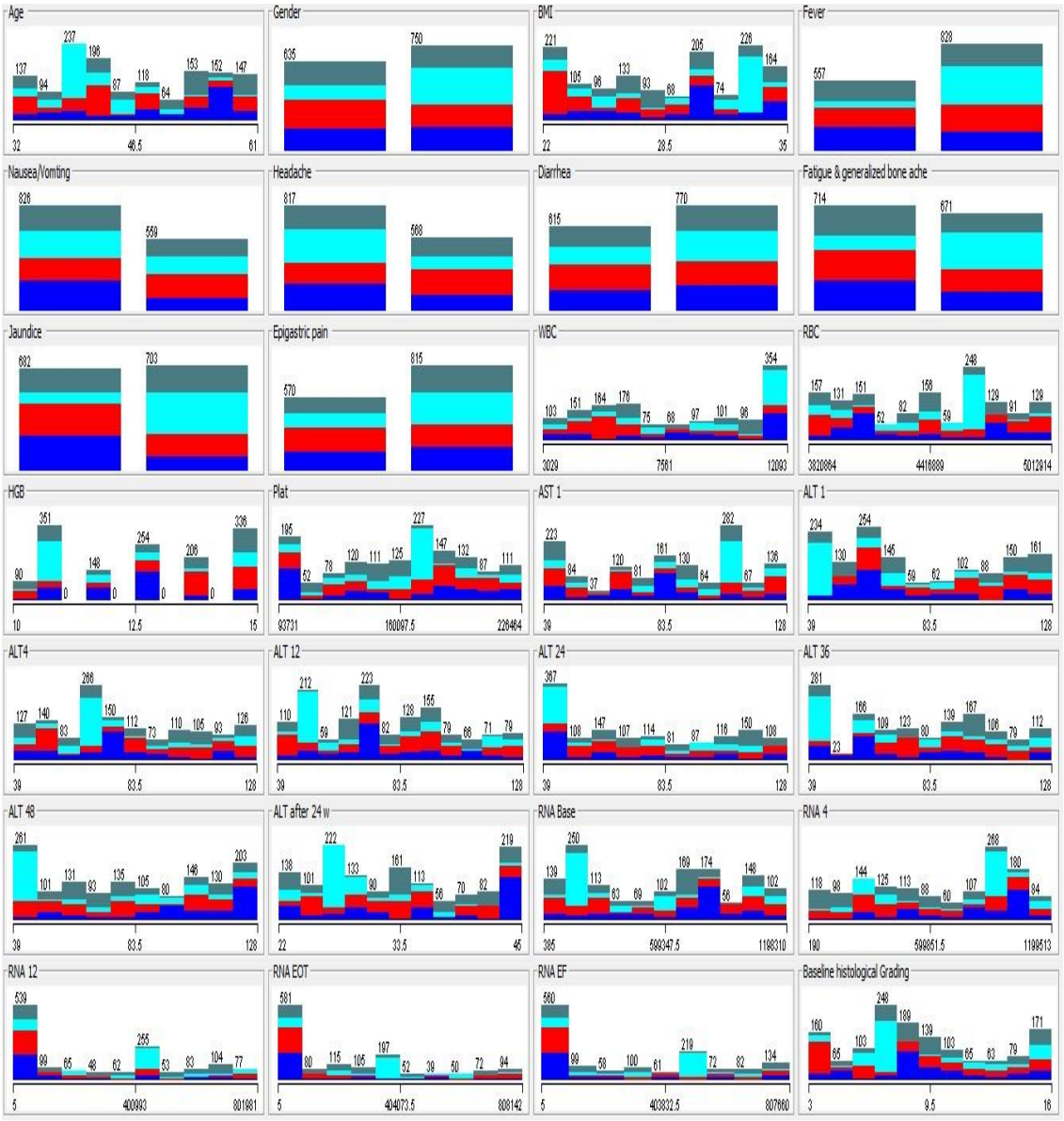


Figure 3.1: Histogram of the Attributes

3.2.3 Data Processing

The ARFF data was imported to WEKA 3.9.4. The histogram of each features was shown in Figure 3.1. From our dataset we have took 80% (1108 instances) of the data for training set and 20 % (277 instances) of data for testing set.

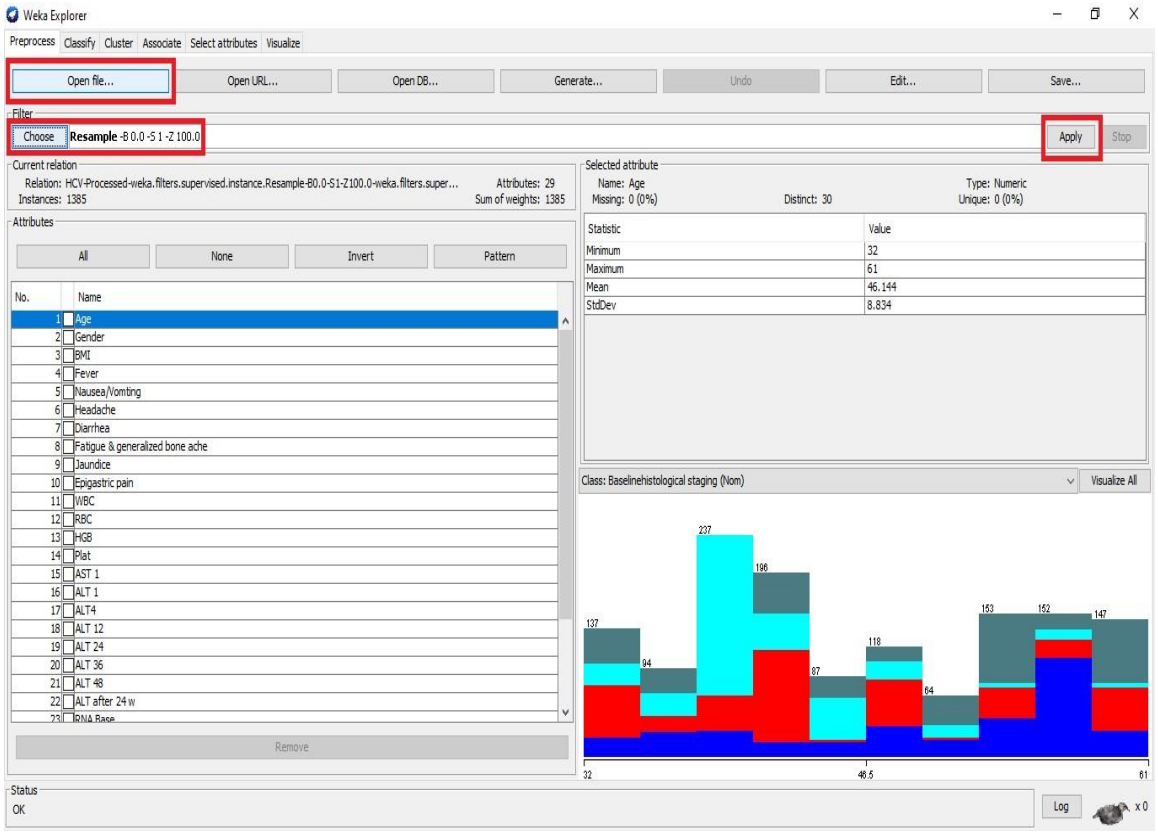


Figure 3.2: Application of the Resample filter

To gain higher accuracy a supervised filter of WEKA tool named Resample was applied shown in Figure 3.2. This function produces subsample to maintain the class distribution in the dataset using either with or without replacement. The filter function can be used to maintain the class distribution in a uniform way.

3.3 Work Flow

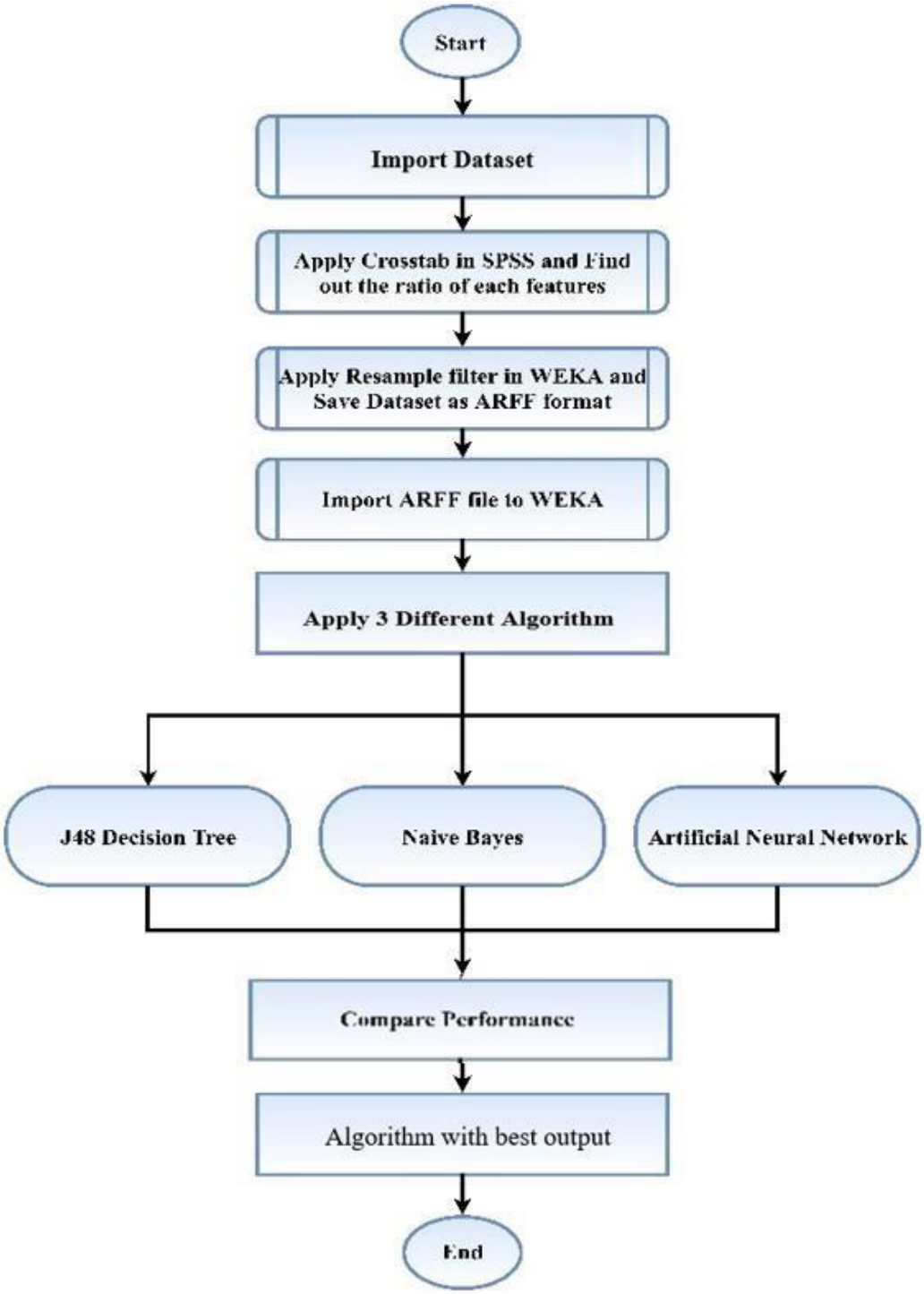


Figure 3.3: Workflow

3.4 Applied Algorithms

3.4.1 J48 Decision Tree

J48 is tree formatted classifier. J48 doesn't learn on its own; We told it first, then it will generate a decision tree to classify. A set of instance's attributes and their corresponding class are used, from that J48 constructs a decision tree that can predict the class for new instances based on their attributes. J48 works faster than most other algorithms. For the value of the given attribute at each point of the tree is a question; depending on those values, the instances get classified. There remain two nodes in a decision tree. They are the decision node and the leaf node. Decisions are made by decision nodes and it contains branches; while Leaf nodes are the product of such decisions as well as they do not have any additional branches. J48 usually imitate human reasoning capacity when taking a verdict.

3.4.2 Artificial Neural Network (ANN)

Artificial Neural Network is applied to allow the device to understand the mapping of attribute values which can be used to identify novel features it works like; the structure of human intellect. ANN is made of interconnected nodes and weighted connections; here nodes in the ANN are considered neurons as equivalent to biological neurons. Based on the particular area of operation, different type network architectures have been developed. There remain basically two kinds of ANN: Feedforward as well as feedback ANN. We have used Multi-Layer Perceptron of ANN; which is used as feed forward pass in ANN. If only one hidden layer is used in the Multi-Layer Perceptron then it can also be called 'vanilla' neural network. Multi-Layer Perceptron consists of input layer; output layer and minimum one hidden layer. It does supervised learning for training; which is called back propagation. It uses sigmoid function as activation function. Multi-Layer Perceptron learns based on the amount of error in the output compared to the expected result so it changes connection weights after each piece of data is processed to acquire result. This algorithm can perform quickly in identify patterns.

3.4.3 Naïve Bayes

Naïve Bayes (NB) is the most direct as well as quick characterization calculation; and can be performed on an enormous piece of information. Naïve Bayes algorithm allows us to predict the class of given set using probability; for that training dataset is needed to be provided to Naïve Bayes in order to construct model. Once the frequency tables are calculated; then it calculates the probabilities for all the classes, and then choosing the highest probability. It is one of the least difficult direct learning calculations. It can perform in high speed on enormous datasets; and it is fast, accurate and reliable algorithm. Despite its simplicity, this algorithm can be surprisingly accurate.

3.5 Research Instruments

For our study purpose we have used SPSS and WEKA.

3.5.1 SPSS

We have used two mainly software for our study; SPSS is one of them. The name SPSS stands for 'Statistical Package for Social Sciences'. We have used really basic function of SPSS to complete our study.

3.5.1.1 Crosstab

In SPSS we have mainly used crosstab function. We have used crosstab to exhibit necessary ratios(percentages) of our study. Also, we have used crosstab to generate our bar charts.

3.5.2 WEKA

We have used WEKA's classifier algorithms to perform our study. WEKA is the short form of 'Waikato Environment for Knowledge Analysis'. WEKA is generally used for building a model. We have used WEKA's resample filter; we got the histogram of all the attributes of our study from WEKA and we have used Multi-Layer Perceptron (MLP) as ANN algorithm, Naïve Bayes algorithm, J48 Algorithm and of WEKA.

CHAPTER 4 EXPERIMENTAL RESULT AND DISCUSSION

4.1 Introduction

In the following chapters we will discuss about the results of the guided experiment. We will research and compare the different classifier accuracy and performance.

4.2 Experimental results and Analysis

Here several attributes were used to predict of the stages liver fibrosis resulted from Hepatitis C Virus. They are- Age, Gender, BMI, Fever, Nausea/Vomiting, Headache, Diarrhea, Fatigue, Jaundice, Epigastric pain, WBC, RBC, HGB, Plat, AST 1, ALT 1, ALT4, ALT 12, ALT 24, ALT 36, ALT 48, RNA Base, RNA 4, RNA 12, RNA EOT, RNA EF, Grading. Relation of some common symptoms of hepatitis c virus with stages of liver fibrosis are shown with bar charts. Besides, the model performance results of the algorithms which were applied in this study to predict of the stages of Hepatitis C Virus were also discussed in this section.

4.2.1 Cross Tab analysis of the symptoms

Figure 4.1 exhibit that the crosstab results among Gender and staging of this exploration.

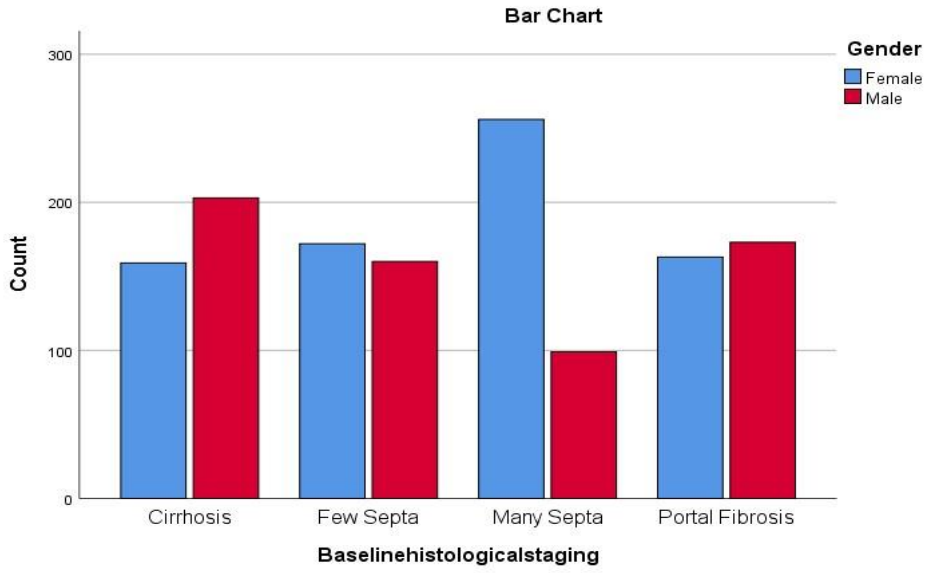


Figure 4.1: Gender with the stages of liver fibrosis caused by HCV

Around 1385 subjects were taken an interest in this investigation and the level of female and male patients were 54.2% and 45.8% individually. In cirrhosis phases of Hepatitis C

Virus 43.9% female patients and 56.1% male patients were acquired. All out number of female patients (51.8%) and male (48.2%) was found in few septa stage of Hepatitis C Virus. A large number of female patients were found in may septa stage of Hepatitis C Virus which amounts were 72.1% and on the other hand only 27.9% male patients were found in this stage. Another stage of Hepatitis C Virus named portal fibrosis confirmed that 48.5% female patients and 51.5% male patients were included in this study. The P value was found 0.001 (<0.05) in this study which was statistically significant.

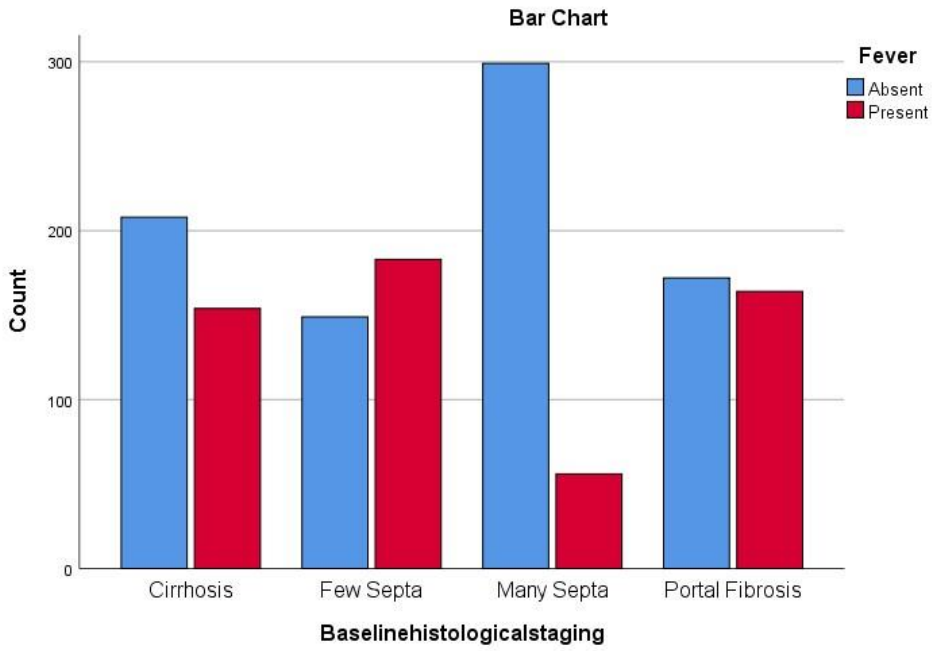


Figure 4.2: Fever conditions of different stages of liver fibrosis caused by HCV

Figure 4.2 indicated the patients Fever states of various phases of Hepatitis C Virus. In cirrhosis phase of hepatitis C virus 57.5% patients were discovered who's the fever was missing and 42.5% patients were discovered who's the fever was available. Another stage of Hepatitis C Virus named few septa found that the fever was absent 44.9% of patients where 55.1% of patients fever condition was present. Moreover 84.2% patients were found who's the fever was missing and 15.8% patients were found who's the fever was present in many septa stage of Hepatitis C Virus. Additionally, in portal fibrosis stage checked 51.2% patients fever condition was missing and 48.8% patients fever condition was

available. The P value was found 0.001 (<0.05) in this study which was statistically significant.

Figure 4.3 showed the patients Nausea or Vomiting states of various phases of Hepatitis C Virus. Nausea or Vomiting symptoms was absent 48.3% of the patients in cirrhosis stage of Hepatitis C Virus where 51.7% of patients has Nausea or vomiting symptoms. In few septa stage 70.2% patients were found whose Nausea or Vomiting symptoms was absent and 29.8% patients had Nausea or Vomiting symptoms. Many septa stage was confirmed the Nausea or Vomiting symptoms and the percentage of absent condition was 60.6% and present condition was 39.4%. Additionally, in portal fibrosis stage checked 60.4% patients Nausea or Vomiting symptom was missing and 39.6% patients Nausea or Vomiting symptom was present. The P value was found 0.001 (<0.05) in this study which was statistically significant.

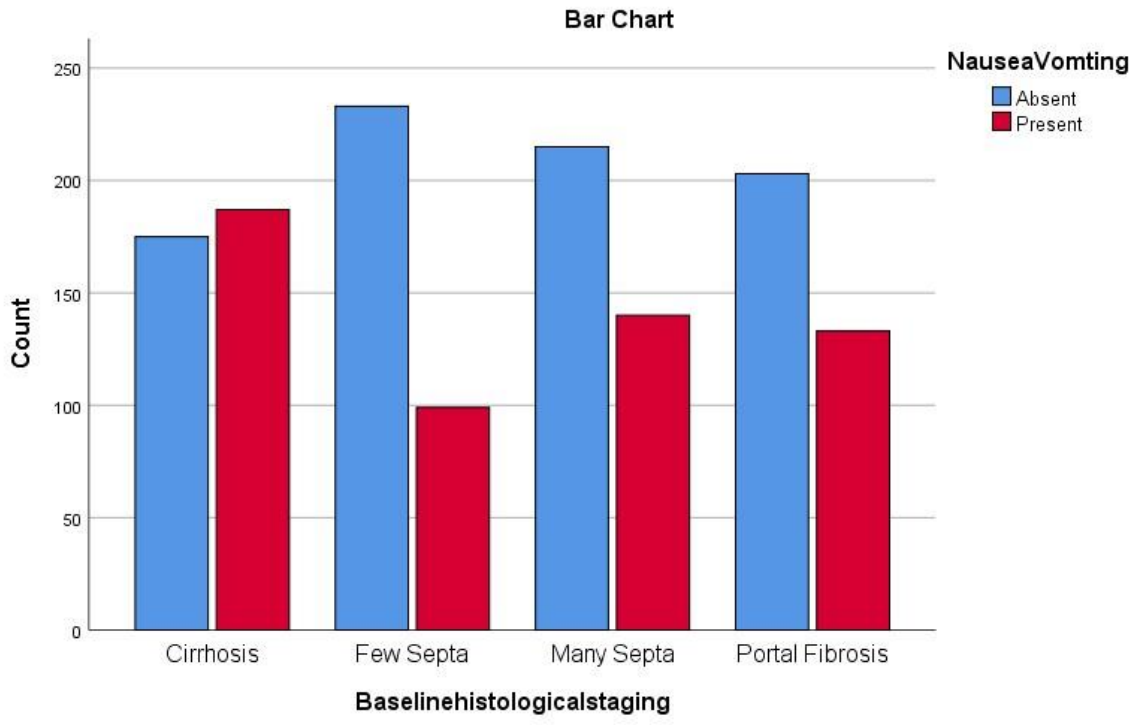


Figure 4.3: Nausea/vomiting conditions of different stages of liver fibrosis caused by HCV

Figure 4.4 showed the patients Headache states of various phases of Hepatitis C Virus. Headache was absent 43.9% of the patients in cirrhosis stage of Hepatitis C Virus where

56.1% of patients had Headache. In few septa stage 63.3% patients were found whose Headache was absent and 36.7% patients had Headache. Many septa stage was confirmed the Headache and the percentage of absent condition was 72.7% and present condition was 27.3%. Additionally, in portal fibrosis stage checked 56.5% patients Headache was missing and 43.5% patients Headache was present. The P value was found 0.001 (<0.05) in this study which was statistically significant.

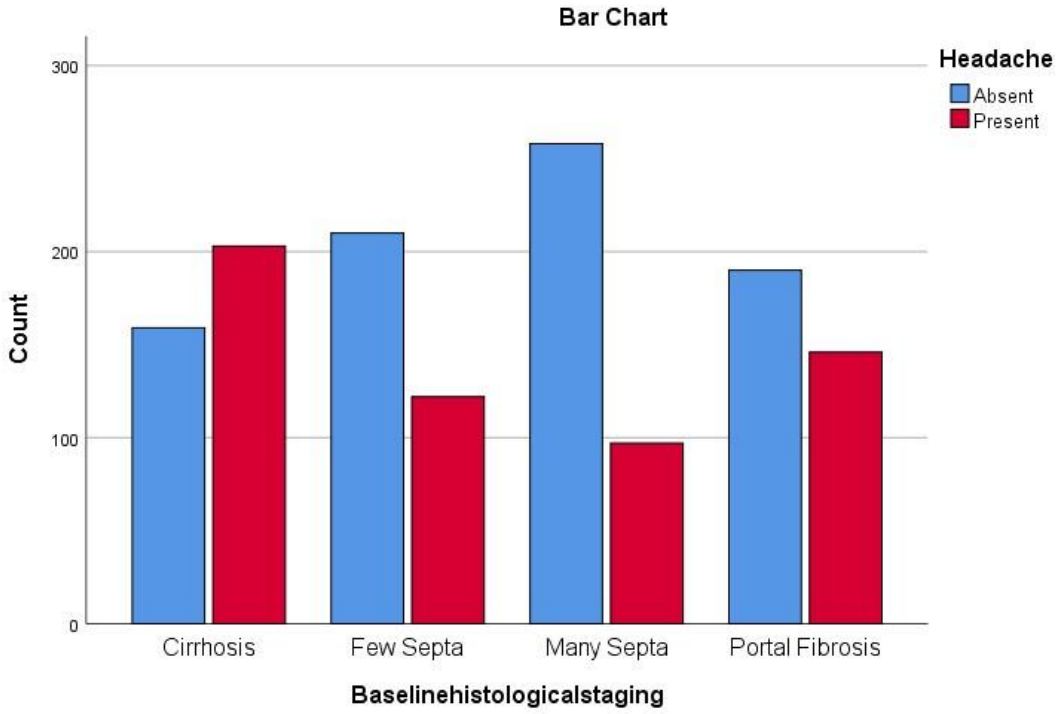


Figure 4.4: Headache conditions of different stages of liver fibrosis caused by HCV

Figure 4.5 indicated the patients Diarrhea states of various phases of Hepatitis C Virus. In cirrhosis phase of Hepatitis C Virus 51.7% patients were discovered who's the Diarrhea was missing and 48.3% patients were discovered who's the Diarrhea was available. Another stage of Hepatitis C Virus named few septa found that the Diarrhea was absent 44.0% of patients where 56.00% of patients Diarrhea condition was present. Moreover 37.2% patients were found who's the Diarrhea was missing and 62.8% patients were found who's the Diarrhea was present in many septa stage of Hepatitis C Virus. Additionally, in portal fibrosis stage checked 44.6% patients Diarrhea condition was missing and 55.4%

patients Diarrhea condition was available. The P value was found 0.001 (<0.05) in this study which was statistically significant.

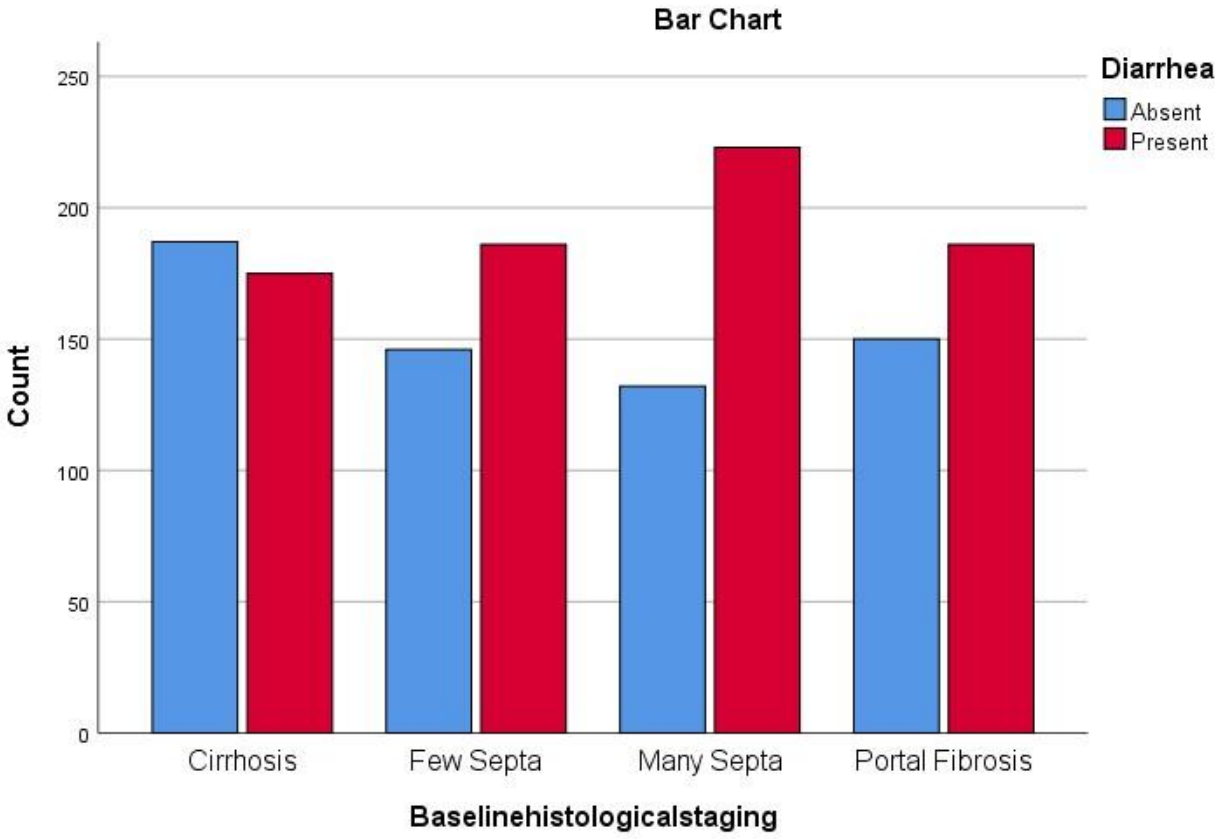


Figure 4.5: Diarrhea conditions of different stages of liver fibrosis caused by HCV

Figure 4.6 showed the patients Jaundice states of various phases of Hepatitis C Virus. Jaundice was absent 40.3% of the patients in cirrhosis stage of Hepatitis C Virus where 59.7% of patients had Jaundice. In few septa stage 29.8% patients were found whose Jaundice was absent and 70.2% patients had Jaundice. Many septa stage was confirmed the Jaundice and the percentage of absent condition was 78.6% and present condition was 21.4%. Additionally, in portal fibrosis stage checked 53.3% patients Jaundice was missing and 46.7% patients Jaundice was present. The P value was found 0.001 (<0.05) in this study which was statistically significant.

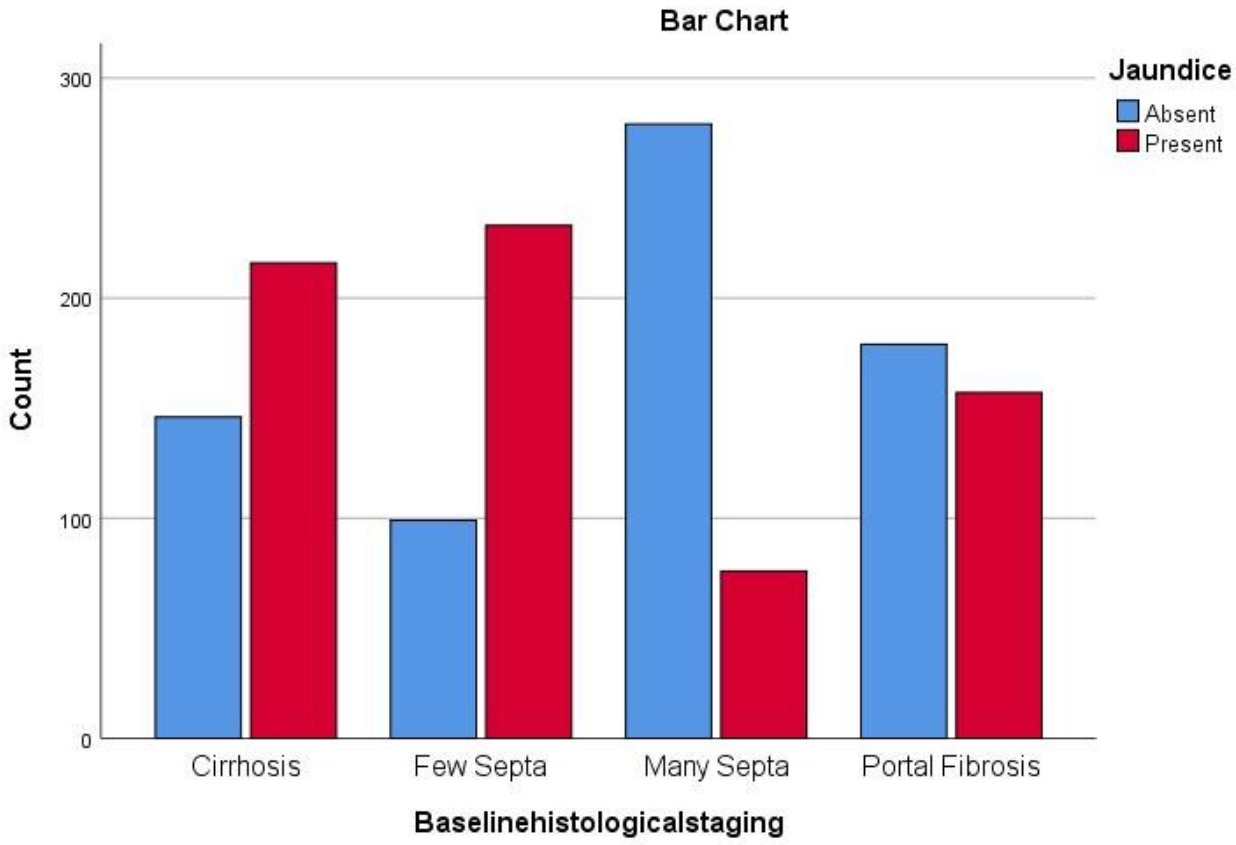


Figure 4.6: Jaundice conditions of different stages of liver fibrosis caused by HCV

4.2.2 Model Performance Result

The accuracy score of applied algorithms was discovered 91.34% for Artificial Neural Network (ANN), 96.03% for J48 and for 94.22% Naïve Bayes.

4.2.3 Confusion matrix

The confusion matrix of the applied algorithms will discuss in this section. It provides us a holistic perspective of how well the classification process works as well as what sorts of mistakes it makes

4.2.3.1 Confusion matrix of J48

The confusion matrix of J48 algorithm showed in figure 4.7 Here Few Septa defined as "A", Cirrhosis defined as "B", Many Septa defined as "C" and Portal Fibrosis defined as "D".

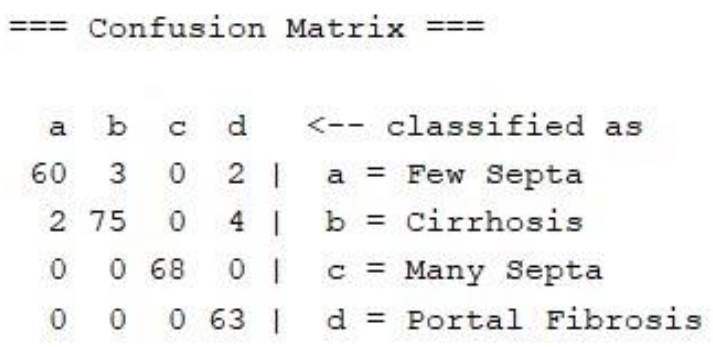


Figure 4.7: Confusion Matrix of J48 Algorithm

From the confusion matrix we found that the Few septa stage was effectively anticipated 60 patients out of 65 patients whose were in Few Septa phase of liver fibrosis. 3 patients were anticipated as Cirrhosis stage and 2 patients as Portal Fibrosis stage however the real yield was in Few Septa stage. Among the 81 patients in Cirrhosis phase of liver fibrosis, the J48 algorithm effectively anticipated 75 patients were in Cirrhosis phase of liver fibrosis. The 2 patients were precited in Few Septa stage and 4 patients were precited in Portal Fibrosis stage however the really the 6 patients were in Cirrhosis phase of Cirrhosis liver fibrosis. The Many septa stage and Portal Fibrosis stage were effectively characterized without error.

4.2.3.2 Confusion matrix of ANN

The confusion matrix of ANN algorithm showed in figure 4.8. Here Few Septa defined as “A”, Cirrhosis defined as “B”, Many Septa defined as “C” and Portal Fibrosis defined as “D”.

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 57  2  3  3 |  a = Few Septa
  6 73  1  1 |  b = Cirrhosis
  1  4 60  3 |  c = Many Septa
  0  0  0 63 |  d = Portal Fibrosis
```

Figure 4.8: Confusion Matrix of ANN Algorithm

From the confusion matrix we found that the Few septa stage was effectively anticipated 57 patients out of 65 patients whose were in Few Septa phase of liver fibrosis. The 3 patients were anticipated as Portal Fibrosis stage however the real yield was in Few Septa stage. The 3 patients were anticipated as Many Septa stage however the real yield was in Few Septa stage. Another 2 patients were anticipated as Cirrhosis stage however the real yield was in Few Septa stage. Among the 81 patients in Cirrhosis phase of liver fibrosis, the ANN algorithm effectively anticipated 73 patients were in Cirrhosis phase of liver fibrosis. The 1 patient were predicted by ANN algorithm was Portal Fibrosis and Many septa respectively where the actual class of those patients was in Cirrhosis stage. Another 6 patients were detected as Few Septa but the actual class was Cirrhosis stage. The Many septa stage was effectively characterized 60 patients among the 68 patients. The 1 and 4 and 3 patients were classified as Few Septa and Cirrhosis and Portal Fibrosis respectively where the actual class was Many Septa stage. The Portal Fibrosis stage had 63 patients and the ANN effectively anticipated all the patients in this stage successfully.

4.2.3.3 Confusion matrix of Naïve Bayes

The confusion matrix of Naïve Bayes algorithm in figure 4.9. Here Few Septa defined as “A”, Cirrhosis defined as “B”, Many Septa defined as “C” and Portal Fibrosis defined as “D”.

```
=== Confusion Matrix ===

 a  b  c  d   <-- classified as
58  5  0  2 |  a = Few Septa
 1 77  0  3 |  b = Cirrhosis
 0  0 65  3 |  c = Many Septa
 0  2  0 61 |  d = Portal Fibrosis
```

Figure 4.9: Confusion Matrix of Naïve Bayes Algorithm

From the confusion matrix we found that the Few septa stage was effectively anticipated 58 patients out of 65 patients whose were in Few Septa phase of liver fibrosis. The 5 patients were anticipated as Cirrhosis stage however the real yield was in Few Septa stage. Another 2 patients were anticipated as Portal Fibrosis stage however the real yield was in Few Septa stage. Among the 81 patients in Cirrhosis phase of liver fibrosis, the Naïve Bayes algorithm effectively anticipated 77 patients were in Cirrhosis phase of liver fibrosis. The 3 patients and 1 patient were predicted by Naïve Bayes algorithm was Portal Fibrosis and Few septa respectively where the actual class of those patients was in Cirrhosis stage. The Many septa stage was effectively characterized 65 patients among the 68 patients. The 3 patients were classified as Portal Fibrosis where the actual class was Many Septa stage. The Portal Fibrosis stage had 63 patients and the Naïve Bayes effectively anticipated 61 patients in this stage successfully. The 2 patients were predicted as Cirrhosis stage where the actual stage was Portal Fibrosis.

4.3 Summary

In our study best accuracy gained from the J48 algorithm among Naïve Bayes, J48 and ANN. The score of J48 is 96.03%.

Table 2: Model Performance Results

Algorithm	ANN	J48	Naive Bayes
Accuracy	93.50%	96.03%	94.22%

Table 3: Comparison between confusion matrix

Number of Patient in each class	J48	ANN	Naive Bayes
65 Patients in Class A	60	57	58
81 Patients in Class B	75	73	77
68 Patients in Class C	68	60	65
63 Patients in Class D	63	63	61

CHAPTER 5 SUMMARY AND CONCLUSION

5.1 Summary of the Study

In our study best accuracy gained from the J48 algorithm and the score is 96.03%; where the rest of two algorithm performed less score then J48 algorithm. Root mean squared is also lowest in J48 algorithm. This study concluded that the J48 algorithm performed best accuracy **(96.03%)**. We have also found that selected symptoms of hepatitis C virus has a statistically significant (P value <0.05) relation with the staging of liver fibrosis.

5.2 Conclusions

We observed that, we have used 3 different algorithms among which J48 had the highest level of accuracy. Although the other algorithms provided very close and accurate result compared to J48 algorithm. So, we can come to a conclusion that J48 decision tree algorithm is the most suitable for staging purpose of liver fibrosis among Naïve Bayes, J48 & ANN. Beside that we found that symptoms of HCV have statistically significant relation with the staging of liver fibrosis.

5.3 Implication for Further Study

For various other diseases and health problem we can use similar techniques. In the future we can do further studies in the field for achieving more clinical accuracy and reliability. In future we can also develop web based or android based or iOS application with permission and supervision from appropriate authorities for public usage of the platform.

REFERENCES:

[1] Vishram Singh, Textbook of Anatomy Abdomen and lower limb, 3rd Edition, Elsevier RELX India Pvt. Ltd., 2019, pp. 105-107.

[2] World Health Organization, available at << https://www.who.int/news-room/fact-sheets/detail/hepatitisc>>, last accessed on 15-11-2020 at 9:00 PM.

[3] Liver Fibrosis, available at << https://www.healthline.com/health/liverfibrosis?fbclid=IwAR2ZDEvv4Izgg7yPepjjss8cik7N1dO2mdLt3ynAaj7mqIRWCpIRSmJk7dE>>, last accessed on 15-11-2020 at 10:00 PM.

[4] Cunningham, Sally Jo, and Geoffrey Holmes. "Developing innovative applications in agriculture using data mining." In The proceedings of the Southeast Asia regional computer confederation conference, pp. 25-
29. 1999.

[5] World Health Organization. Secretariat. Viral hepatitis. Sixty-Third World Health Assembly A63/15. Provisional agenda item 11.12. Geneva: World Health Organization; 2010.

[6] Al-Mahtab, M., Past, Present, and Future of Viral Hepatitis in Bangladesh. Euroasian Journal of HepatoGastroenterology, 6(1), pp.43-44,2016.

[7] K.Park, Park's Textbook of Preventive and Social Medicine, 25th edition, M/s Banarsidas Bhanot, 2019, pp. 230-240.

[8] Vinay Kumar, Abul K. Abbas, Jon. C. Aster, Robbins and Cotran Pathologic Basis of Disease Systemic Pathology, 1st Bangladesh Edition, Elsevier RELX India Pvt. Ltd., 2017, pp. 822-834.

[9] The AIDS InfoNet, available at << http://www.aidsinfonet.org/fact_sheets/view/670>>, last accessed on 15-11-2020 at 10:15 PM.

[10] Healthline, available at << https://www.healthline.com/health/stage-4-hepatitis-c>>, last accessed on 15-11-2020 at 10:30 PM.

[11] MDPI, available at << https://www.mdpi.com/1422-0067/21/11/4039 >>, last accessed on 16-11-2020 at 8:00 PM.

[12] MAYO CLINIC, available at << https://www.mayoclinic.org/diseases-conditions/cirrhosis/symptomscauses/syc-20351487?fbclid=IwAR0JYAKmX7ydA5ruDcICDmahjc35IZZQrmuZlTXxMsXu61G4tPojCJjuzU0#:~:text=Cirrhosis%20is%20a%20late%20stage,it%20tries%20to%20repair%20itself>>, last accessed on 16-112020 at 8:30 PM.

[13] Yarasuri, Vedha Krishna, Gowtham Kishore Indukuri, and Aswathy K. Nair. "Prediction of Hepatitis Disease Using Machine Learning Technique." In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 265-269. IEEE, 2019.

[14] Ahmed A. A. Radwan, Tarekabd-El-Hafeez & Heba Mamdouh, "An analysis of hepatitis c virus prediction using different data mining techniques," *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)*, vol. 3, no. 4, pp. 209-220, 2013.

[15] Nandipati, Satish CR, Chew XinYing, and Khaw Khai Wah. "Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques." Applications of Modelling and Simulation 4 (2020): 89-100.

[16] Yada, Norihisa, Masatoshi Kudo, Norifumi Kawada, Shuichi Sato, Yukio Osaki, Akihisa Ishikawa, Hisaaki Miyoshi et al. "Noninvasive diagnosis of liver fibrosis: utility of data mining of both ultrasound elastography and serological findings to construct a decision tree." Oncology 87, no. Suppl. 1 (2014): 63-72. [17] Ayeldeen, Heba, Olfat Shaker, Ghada Ayeldeen, and Khaled M. Anwar. "Prediction of liver fibrosis stages by machine learning model: A decision tree approach." In 2015 Third World Conference on Complex Systems (WCCS), pp. 1-6. IEEE, 2015.

[18] Keltch, Brian, Yuan Lin, and Coskun Bayrak. "Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients." Journal of medical systems 38, no. 8 (2014): 60.

[19] Barakat, Nahla H., Sana H. Barakat, and Nadia Ahmed. "Prediction and Staging of Hepatic Fibrosis in Children with Hepatitis C Virus: A Machine Learning Approach." Healthcare Informatics Research 25, no.

3 (2019): 173-181.