A Dissertation work entitled on

Classification and Prediction of different types of liver diseases using Machine Learning Techniques

A Project Report submitted in partial fulfillment of the requirements for the award of the degree of

Master of Technology

in

Software Engineering

By

Hajera Subhani

160617742305

Under the guidance of

Dr. Srinivasu Badugu

Associate Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,

Stanley College of Engineering and Technology for Women,

(Affiliated to Osmania University)

Chapel Road, Abids, Hyderabad

2018-2019

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,

Stanley College of Engineering and Technology for Women,

(Affiliated to Osmania University)

CERTIFICATE

This is to certify that the project entitled **“Classification and Prediction of different types of liver diseases using Machine Learning Techniques”** submitted by **Hajera Subhani (160617742305),** a student of Department of Computer Science Engineering, Stanley college of engineering and technology for women’s in partial fulfillment of the requirements for the award of the degree of Master of Technology with Software Engineering as specialization is a record of bonafide work carried out by her during academic year 2018-2019.

Signature of the guide	Signature of the Head of the Dept.
Dr. Srinivasu Badugu	**Dr. B.V Ramana Murthy**
Associate Professor,	**PROFESSOR & HOD,**
Dept. of CSE.	**Dept. of CSE.**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,

Stanley College of Engineering and Technology for Women,

(Affiliated to Osmania University)

DECLARATION

I declare that the work reported in the thesis entitled **"Classification and Prediction of different types of liver diseases using Machine Learning Techniques"** is record of the work done by me in the Department of Computer Science and Engineering and Technology for women, Hyderabad.

No part of the thesis is copied from books/journals internet and wherever referred the same has been duly acknowledge in the text. The reported data is based on the project work done entirely by me and not copied from any other source.

Hajera Subhani

(160617742305)

To Daddy

Every Father is a super hero for their daughter and if her father is facing any problem then daughter cannot bear that. This is the work dedicated to you.

To Ami

Ami you're my mother in the eyes of this world but for me you're everything for me because you encourage me to fulfill every dream which I see with my open eyes. They are no words to explain my feelings about you-Ami.

"Worry weighs a person down; an encouraging word cheers a person up."

ACKNOWLEDGEMENT

It is with a sense of gratitude and appreciation that I feel to acknowledge any well-wishers for their king support and encouragement during the completion of the project.

I thank **Dr. Satya Prasad Lanka**, the Principal, Stanley College of Engineering and Technology for women, for his timely cooperation and for providing me all the required facilities to complete the project successfully.

I would extremely grateful to **Dr. B V Ramana murthy** Head of the department for providing excellent computing facilities and such a nice atmosphere for completing my project.

I would like to express my heartfelt gratitude to **Dr. Srinivasu Badugu**, Coordinator of M.Tech in the Computer science Engineering department and my internal supervisor, for encouraging and guiding me through the project. I am highly indebted to him for his guidance and constant supervision regarding the project as well as for providing valuable suggestions and continuous support given to me for completing the project successfully and all those who helped me directly or indirectly during the course of project. I sincerely thank all of them.

Hajera Subhani

(160617742305)

ABSTRACT

CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE NO.	CAPTION	PAGE NO.

CHAPTER 1 INTRODUCTION

Diagnosis of liver infection at preliminary stage is important for better treatment of the patients who are suffer from liver diseases. We utilize machine learning approaches for classification of patients who are suffering from liver disease and for predicting different types of liver diseases they are suffering from. ML approach such as SVM and KNN are used for classification. This project is to somewhat reduce the time delay caused in the treatment due to the unnecessary back and forth shuttling between the hospital and the pathology lab. Delayed diagnosis of diseases is a fundamental problem due to a shortage of medical professionals. A typical scenario, prevalent mostly in rural and somewhat in urban areas. Automatic classification tools may reduce the burden on doctors.

The proposed system will be classifying the patients who are suffering from liver diseases using Machine Learning Classification techniques and further classifying the type of diseases the patient is suffering from.

Classification techniques are very popular in various automatic medical diagnosis tools Early identification of cancer has been often vital for the survival of the patients.

Classification is a supervised **learning** approach in which the computer program learns from the data input given to it and then uses this **learning** to **classify** new observation. [23]

1.1 Liver

Liver is one of the vital organs of the body that sits on the right side of the stomach. It is reddish-brown, rubbery organ that is well protected beneath the rib cage. Liver is the largest internal organ in the human body and is responsible for many critical functions within the body.[3] Damage or diseased liver result in the loss of those functions and significant damage to the body.Usually,75% of the liver need to be damaged or been affected to decrease the functionality of liver.

The two large sections of the liver, called as right and left lobe, along with pancreas and gallbladder help to digest, absorb and process food. Liver secretes bile which is essential for digestion and makes protein which is essential for clotting and other functions. The blood coming from the digestive tract reaches liver for removing toxins and metabolizing drugs before it is supplied to the body.

1.1.1 Functions of liver

- Produces bile a bio chemicals substance needed for digestion.
- Maintains proper levels of glucose in the blood.
- Converts ammonia to urea which is vital in metabolism.
- Stores and Destroys old RBC.

- Stores Vitamins (A,D,K & B12) and iron (also stores minerals).
- Stores and releases glucose.
- Controls metabolism.
- Detoxifies the blood (filters harmful substances form the blood, such as alcohol).
- Protein synthesis.
- Produces 80% of your body's cholesterol.
- Stores glucose
- Synthesizing plasma protein
- The production of hormones
- Produces urea
- Controls metabolism [18]

1.1.2 Signs and symptoms of liver disorders:

Most types of liver diseases do not show any signs and symptoms in the early stages. The symptoms are visible when the liver is already damaged or scarred. Some of the common signs and symptoms of liver disorders are:

- Chronic tiredness
- Decreased appetite (hunger)
- Discoloration of the urine
- Easy bruising
- Leg and ankle swelling
- Nausea or vomiting
- Itchy skin
- Swelling and pain in the abdomen
- Tar coloured or clay-like pale stools
- Yellowish coloration of eyes and skin; known as jaundice[18]

Complications of liver disorders:

The complications and outcomes of liver disease depend on the underlying cause. However, a long-standing and untreated liver disease can develop into a life threatening liver failure. Some of the complications of liver disorders include:

- Swelling (edema) and accumulation of fluids in the abdomen (ascites)
- Bruising and bleeding in liver

- Portal hypertension (high pressure in portal veins which supply to liver from GI tract, pancreas, gallbladder and spleen)
- Enlarged spleen (splenomegaly)
- Jaundice
- Hepatic encephalopathy
- Insulin resistance and type 2 diabetes
- Liver cancer (hepatocellular carcinoma)
- Worsening immunity (Immune system dysfunction) and increased risk for infections. [18]

1.1.3 Causes of liver disease:

People with the following risk factors are more prone to develop diseases

- Excess alcohol intake
- Shared needles for injections (common in drug abusers)
- Body piercing and tattooing with unsterilized needles
- Exposure to an infected person’s blood and body fluids
- Unprotected sex
- Exposure to harmful chemicals or toxins
- Underlying medical conditions like diabetes, obesity etc.[18]

1.1.4 Liver diseases

Liver disease or hepatic disease is any condition that damages the liver and poorly affects its functioning. Longstanding or untreated liver disease can cause serious, irreversible damage to the liver. Liver damage may be caused in several ways:

- Inflammation of liver as seen in hepatitis.
- Obstruction to bile flow as seen in cholestasis.
- Deposition of excess cholesterol as seen in steatosis.
- Deposition of excess fat as seen in fatty liver and liver enlargement.
- Scarring and damage to liver tissue as seen in fibrosis and cirrhosis.
- Cancerous cells infiltrating as seen in hepatocellular carcinoma. [18]

Types of liver diseases

According to NHS, there are over 100 types of liver diseases. Some of the common types of liver diseases are –

- **Viral (Infectious) hepatitis:** Hepatitis literally means inflammation (swelling and reddening) of the liver. It is caused by the hepatitis viruses, Hepatitis A, B, C, D and E. Hepatitis A and B are preventable by vaccines.

Hepatitis A: is often mild and patients usually make a full recovery in 2 months without intervention. There are safe and effective vaccinations against Hepatitis A virus (HAV).

Hepatitis B: The liver damage may progress from inflammation to cirrhosis and hepatocellular carcinoma. There is no cure for HBV. However, protective vaccines are available.

Hepatitis C: Similar to HBV, hepatitis C can cause severe liver damage, cirrhosis and lead to hepatocellular carcinoma (a form of liver cancer).

Liver cysts: Liver cysts are fluid-filled spaces in the liver. They are usually symptomless and do not need treatment. Larger cysts may cause pain and discomfort which require drainage and removal of the cysts.

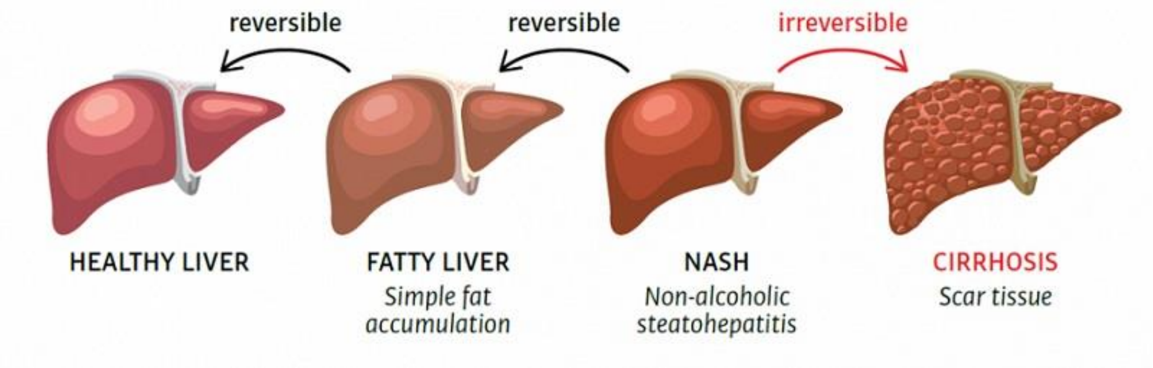
- **Liver Cancer:** Liver cancer is the growth and spread of abnormal, unhealthy cells in the liver. Hepatocellular carcinoma is the most common type of liver cancer. Cirrhosis and Hepatitis B are the leading risk factors for liver cancer.

Hereditary diseases: Some of the genetic liver diseases include,

Wilson disease is the condition of excess copper accumulation in vital organs.

Hemochromatosis is the condition of excess iron retention in the body.

Cirrhosis: Cirrhosis is the scarring of the liver where the soft healthy tissues are replaced with hard scar tissue. Cirrhosis may result from longstanding inflammation from infections, heart disease, or continuous injury.



- **Fatty liver**, there is excess fat stored in the liver. It can either be an alcohol-related fatty liver disease (ALD) or Nonalcoholic Fatty Liver Disease (NAFLD).

Alcohol-related (fatty) liver disease is marked by alcoholic hepatitis and alcoholic cirrhosis.

Non-alcoholic fatty liver disease (NAFLD) is either simple fatty liver or Non-Alcoholic Steato Hepatitis (NASH). In simple fatty liver, the liver cells are inflamed. However, in NASH, the liver cells are further damaged with fibrosis, cirrhosis and even cancer of the liver. [18]

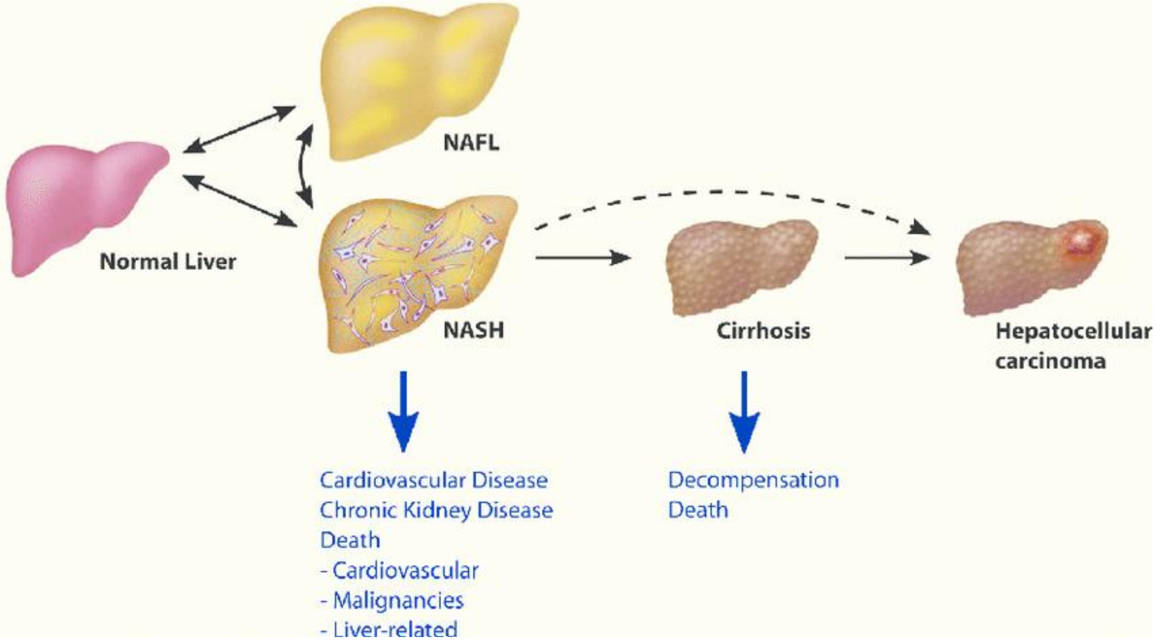


Fig: Non-alcoholic liver disease (NAFLD) spectrum. **[24]**

Stages of liver diseases:

There are several conditions that are diagnosed with liver disease. However, the damage to the liver follows a consistent pattern from the initial stages to advanced stages of the disease.

Stage 1 – Inflammation: Irrespective of the cause of the liver disease, the liver and the liver ducts get inflamed (swollen, reddened) causing abdominal pain. If left untreated, inflammation can cause further damage to the tissues. Inflammation of liver is often treated completely.

Stage 2 – Fibrosis: In many cases, liver diseases may not be diagnosed until stage 2. Fibrosis of

liver is marked by scarring in the tissues which may affect blood flow to liver and liver functions. With treatment, the scarring is healed and further damage is prevented.

Stage 3 – Cirrhosis: Cirrhosis of the liver is a chronic (longstanding) condition marked by permanent scarring that obstructs blood flow to the liver. The most common causes of cirrhosis in the US are chronic hepatitis C infection and alcoholic liver disease. This stage of liver disease is serious and the treatment must begin immediately to halt the progress of the liver disease and the damage. Liver cirrhosis causes decompensation of the liver and can cause serious symptoms and comorbid conditions which need to be managed with prompt care. It is important to protect the healthy tissues that are left to retain the liver functions intact.

Some of the alarming changes in the decompensated cirrhosis (liver disease) include portal hypertension, esophageal varices (dilated, ballooned veins), ascites (accumulation of fluid in the abdominal cavity) and gastrointestinal bleeding. The patient may experience extreme fatigue, confusion, personality changes, extreme sleepiness, reduced urination (an indication of kidney failure), high fever (an indication of possible abdominal infection), swelling in the extremities, wasting of muscles in the extremities, hand tremors, shortness of breath, pale/yellow skin, weight loss, loss of appetite.

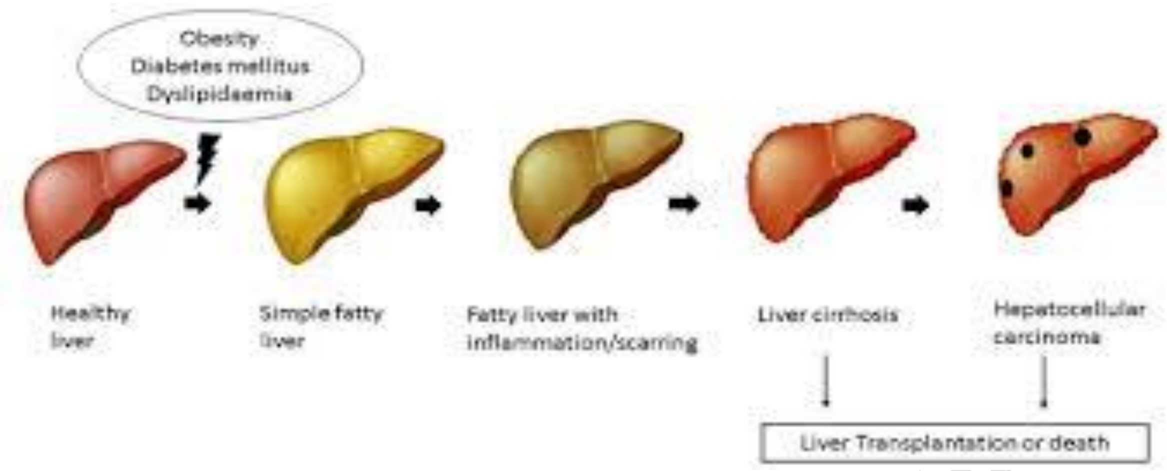
In the presence of cirrhosis, hepatocellular carcinoma (HCC) accounts for approximately 75% of all liver tumors. The most important risk factors for development of HCC are cirrhosis of any cause, hepatitis B virus infection, and hepatitis C virus infection with advanced fibrosis. [13]

Stage 4 – Liver failure: Liver failure, also known as liver insufficiency, is the condition wherein the normal functions of the liver begin to fail. Large part of the liver is affected by irreparable damage and thus the liver fails to perform the routine activities. Based on the causative factors, liver failure may either be acute (rapid development usually in patients without known prior liver disease), chronic (slow progress due to long-term, excessive alcohol intake, hepatitis B, C etc.), or **acute-on-chronic liver failure** (ACLF).

Acute liver failure can be life-threatening and needs immediate medical attention. Usually, it develops within a few days due to over dosage or poisoning from medications such as paracetamol, infections (hepatitis B or C), acute fatty liver of pregnancy, etc.**Chronic liver failure** develops slowly over many years due to long-term exposure to excessive alcohol, infections.

Acute-on-chronic liver failure occurs as a result of alcohol misuse or infection.Patients with ACLF can be critically ill and require intensive care and sometimes liver transplants.

End-stage liver disease (ESLD): Extensive damage most commonly as a result of cirrhosis or long-term infections results in chronic liver failure. Patients with ESLD show symptoms and complications that affect survival and quality of life. [18]

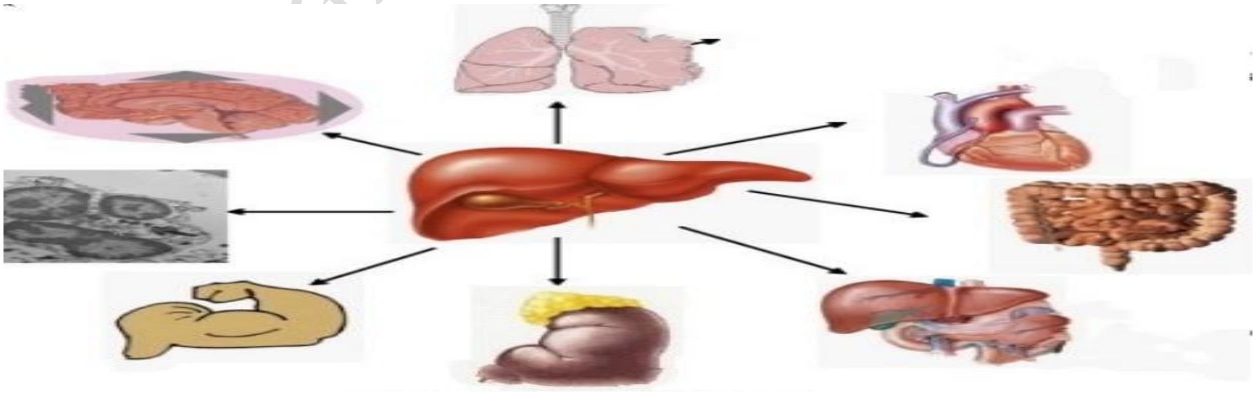


Liver diseases diagnosed:

It is important to determine the underlying cause of liver disease. Your physician or gastroenterologist shall determine the diagnosis on the basis of:

- Medical and personal history
- Physical examination
- Blood tests
- Liver function tests
- Complete blood count (CBC); blood clotting tests; lipase; electrolytes and creatinine, etc.
- Imaging studies (CT scan, MRI and ultrasound)
- Liver biopsy [18]

1.1.5 Effect of liver disease on different organs



Lungs: Liver disease can “portal hypertension,” **meaning** increased blood pressure in the veins that enter the **liver**. As a result, the blood vessels of the **lung** are exposed to possible toxic substances and this can **damage** the small arteries of the **lungs**, causing pulmonary arterial hypertension (PAH).

Heart: The **liver diseases** affecting the **heart** include complications of **cirrhosis** such as hepatopulmonary syndrome, portopulmonary hypertension, pericardial effusion, and cirrhotic cardiomyopathy as well as noncirrhotic **cardiac disorders** such as high-output **failure** caused by intrahepatic arteriovenous fistulae.

large and small intestine: Inflammation and HIV causes **large and small intestine** failure in the inflammation liver does not produces the bile so there is the problem in the digestion of food and thus it leads to the development of the cancer in the intestines.

Pancreases: Liver Produces the bile which is passed to intestine if it is not working properly(blocked) than it means tumors in bile ducts and pancreases is developed .

Kidney: Renal failure(for example paracetomol overdose) results in liver failure which further leads to kidney failure.

Arms: Muscle depletion is a common feature of chronic liver disease found in 40% of patients with cirrhosis

Skin: Hepatitis c affects skin such as rashes and itchy spots

Brain: Researchers believe that high blood pressure of the main vein of the liver patient (hyper tension) results in blood bypyassing the liver. Liver disease affects the brain in other disorders such as reye syndrome and metabolic disorder. [18]

1.1.6 Liver diseases are prevented:

Liver disorders may be prevented with a few lifestyle changes and precautions.
Avoid heavy alcohol intake
Avoid indulging in risky practices like intravenous drug abuse, unprotected sex etc.
In case of accidental exposure to body fluids of a Hepatitis infected person, talk to your doctor or a gastroenterologist about getting vaccinated for Hepatitis B.
Use the medications wisely, whether prescribed by your doctor or bought over-the-counter.
Adopt safe practices while using sprays like insecticides, fungicides, toxic chemicals, paints etc. Maintain a healthy body weight. [18]

1.2 Problem Statement and Description

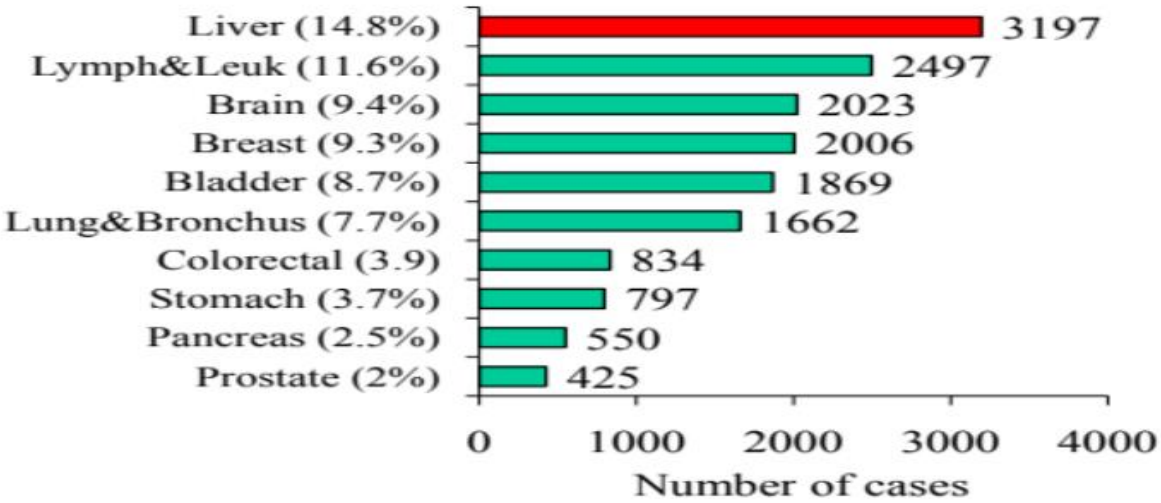


Figure . Egypt Mortality Statistics, Most common sites (The Cancer Database, 2001)

liver cancer the most common cancer in many parts of the world.

- According to WHO Cancer is the second leading cause of death globally, and is responsible for an estimated 9.6 million deaths in 2018. The majority of cancer deaths is caused by liver, lung, breast, colorectal and stomach cancer . Globally, about 1 in 6 deaths is due to cancer.
- Liver Cancer is a leading cause of death worldwide, accounting for an estimated deaths 782 000 deaths in 2018.
- According to times of India, in India one in five persons is affected by liver disease. It is expected by the experts that by 2025 India may become the ‘World Capital for Liver cancer’.
- According to the American Cancer Society, hepatitis C is the most common cause of liver cancer in the U.S. People with both hepatitis B or C have a significantly higher risk of developing liver cancer than other healthy individuals, as both forms of the disease can result in cirrhosis.
- Early identification of the liver cancer will play vital role for the survival of the patients. Automatic tools may reduce the burden on doctors.

1.3 Motivation:

The need of this project work of classification and prediction of different liver diseases is to protect our world from the effect of different liver diseases which are affecting large number of our world’s population from a decade.

Delayed diagnosis of diseases is a fundamental problem due to a shortage of medical professionals. A typical scenario, prevalent mostly in rural and somewhat in urban areas. Automatic classification tools may reduce the burden on doctors and also may reduce the death percentage due to different liver diseases.

Actually it is correct to say that, if you are suffering from any problem you get to know how tough time others goes through in the same problem that you are suffering from. My father who is suffering from non-alcoholic fatty liver and thus I got to know how tough time my father was suffering from than I realized and started reading about liver and had done a lot of research work.

I got motivated to develop a model which can classify as well as predict the different types of liver diseases which our world is suffering from every year by seeing the millions of deaths which are caused due to various liver diseases around the world to make such model which can classify and predict different liver disease.

In this work, person with the liver disease and the person who is not suffering from the liver disease are classified. The person who is suffering from the liver disease is predicted with the type of liver disease he or she is suffering from.

1.4 Aim and scope

Aim of the project: The aim of this project is to build a model which can predict the patients with the liver disease and also predict the various types of liver diseases our world is suffering from. So, that early identification of the liver cancer will play vital role for the survival of the patients. Automatic tools may reduce the burden on doctors.

Aim to predict different types of liver disease the person is suffering from after classification.

Aim to predict the % of damage to the liver of the person who is suffering from different liver diseases.

Aim to predict Cirrhosis which is end stage of liver cancer.

Aim to guide the patients who are suffering from the various liver diseases.

Scope of the project: The scope of the project is to improve the prediction system of different liver diseases from manual to automatic system by using different datasets for training the model and using different machine learning algorithm for classification such as SVM and KNN.

Based on confusion matrix the performance of different algorithm is compared and the algorithm who is having the highest accuracy and performance rate is selected for classification.

1.5 Background and basics

1.5.1. Machine learning

Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed.

The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output value within an acceptable range.

Machine learning refers to a system capable of acquiring and integrating the knowledge automatically. The capability of the systems to learn from experience.

Machine learning is a science of getting computers to act by feeding them data and letting them learn a few tricks on their own, without being explicitly programmed to do so

The key to machine learning is the data. Machines learn just like us humans. We humans need to collect information and data to learn, similarly, machines must also be fed data in order to learn and make decisions.

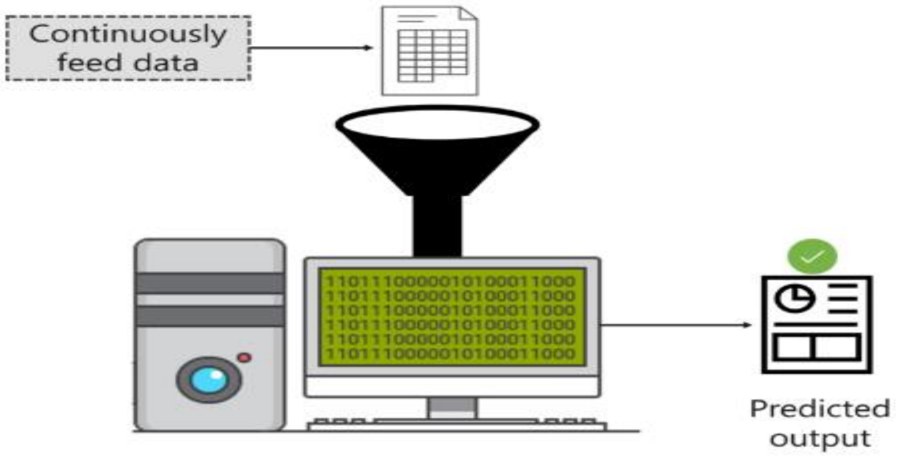


Fig: machine learning[19]

1.5.2 KNN Algorithm

The k-nearest neighbors (**KNN**) **algorithm** is a simple, easy-to-implement supervised **machine learning algorithm** that can be used to solve both classification and regression problems. **The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. Thus it is named as KNN** (K-Nearest Neighbors)**.** [21]

What is KNN Algorithm?

K nearest neighbors or KNN Algorithm is a simple algorithm which uses the entire dataset in its training phase. Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances and the data with the most similar instance is finally returned as the prediction. [22]

KNN is often used in search applications where you are looking for **similar** items, like **find items similar to this one.**

Algorithm suggests that if ***you're similar to your neighbours, then you are one of them***. For example, if apple looks more similar to peach, pear, and cherry (fruits) than monkey, cat or a rat (animals), then most likely apple is a fruit.

How does a KNN Algorithm work?

The k-nearest neighbors algorithm uses a very simple approach to perform classification. When tested with a new example, it looks through the training data and finds the k training examples that are closest to the new example. It then assigns the most common class label (among those k-training examples) to the test example.

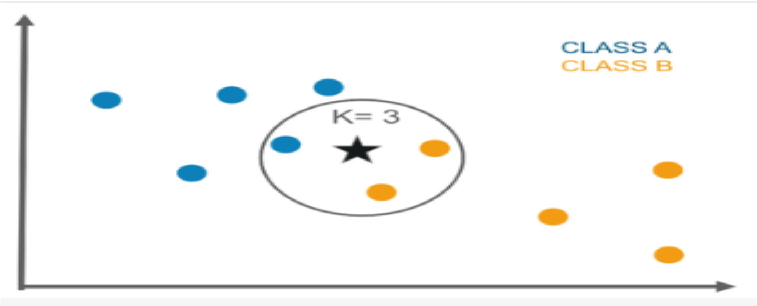
What does 'k' in KNN Algorithm represent?

"k in kNN algorithm represents the number of nearest neighbor points which are voting for the new test data's class".

If k=1, then test examples are given the same label as the closest example in the training set.

If k=3, the labels of the three closest classes are checked and the most common (i.e., occurring at least twice) label is assigned, and so on for larger k's. [22]

Using the k-nearest neighbor algorithm we fit the historical data (or train the model) and predict the future. [19]



How do we choose the factor K?

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Euclidean distance measures is only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

A few Applications and Examples of KNN

- Credit ratings
- Should the bank give a loan to an individual? Would an individual default on his or her loan? Is that person closer in characteristics to people who defaulted or did not default on their loans?
- In political science—classing a potential voter to a “will vote” or “will not vote”, or to “vote Democrat” or “vote Republican”.
- More advance examples could include handwriting detection (like OCR), image recognition and even video recognition.

Some pros and cons of KNN

Pros:

- No assumptions about data—useful, for example, for nonlinear data

- Simple algorithm—to explain and understand/interpret
- High accuracy (relatively)—it is pretty high but not competitive in comparison to better supervised learning models
- Versatile—useful for classification or regression

Cons:

- Computationally expensive—because the algorithm stores all of the training data
- High memory requirement
- Stores all (or almost all) of the training data
- Prediction stage might be slow (with big N)
- Sensitive to irrelevant features and the scale of the data

Summary of KNN

The algorithm can be summarized as:

A positive integer k is specified, along with a new sample

1. We select the k entries in our database which are closest to the new sample
2. We find the most common classification of these entries
3. This is the classification we give to the new sample

A few other features of KNN:

- KNN stores the entire training dataset which it uses as its representation.
- KNN does not learn any model.
- KNN makes predictions just-in-time by calculating the similarity between an input sample and each training instance.

Breaking it Down – Pseudo Code of KNN

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points

1. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other metrics that can be used are Chebyshev, cosine, etc.
2. Sort the calculated distances in ascending order based on distance values
3. Get top k rows from the sorted array
4. Get the most frequent class of these rows
5. Return the predicted class

1.5.3 SVM (Support Vector Machine)

SVM (Support Vector Machine) is a supervised machine learning algorithm which is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyperplane which acts like a decision boundary between the various classes.

SVM can be used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data.

Features of SVM:

1. SVM is a supervised learning algorithm. This means that SVM trains on a set of labeled data. SVM studies the labeled training data and then classifies any new input data depending on what it learned in the training phase.
2. A main advantage of SVM is that it can be used for both classification and regression problems. Though SVM is mainly known for classification, the SVR (Support Vector Regressor) is used for regression problems.
3. SVM can be used for classifying non-linear data by using the kernel trick. The kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data. After which you can easily draw a hyperplane between the various classes of data.

How Does SVM Work?

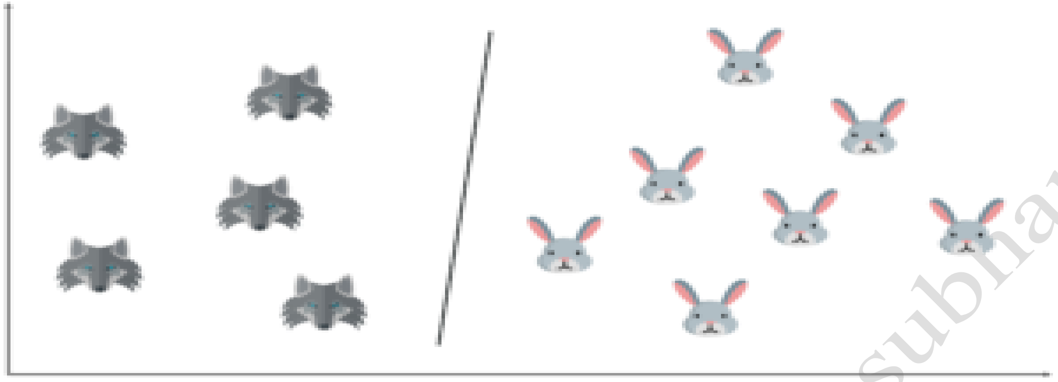
In order to understand how SVM works let's consider a scenario.

For a second, pretend you own a farm and you have a problem–you need to set up a fence to protect your rabbits from a pack of wolves. But where do you build your fence?

How does SVM work? – Support Vector Machine

One way to get around the problem is to build a classifier based on the position of the rabbits and wolves in your pasture.

So if I do that, and try to draw a decision boundary between the rabbits and the wolves, it looks something like this. Now you can clearly build a fence along this line.



How does SVM work? – Support Vector Machine

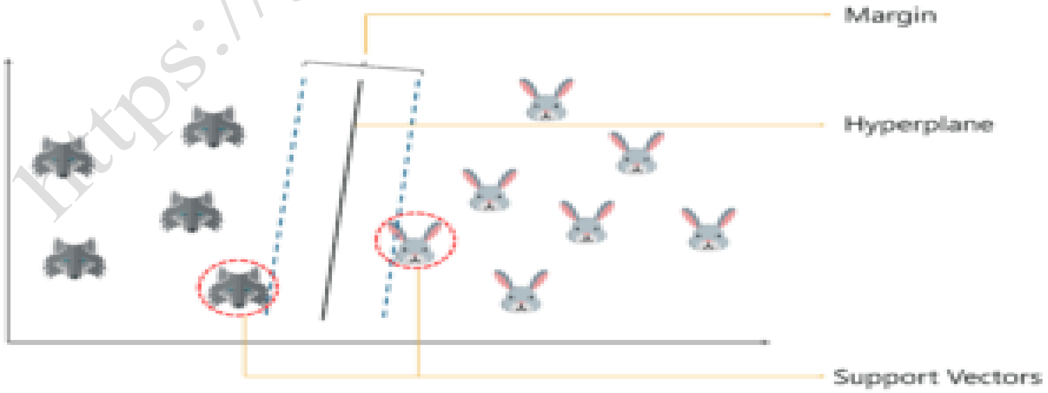
In simple terms, this is exactly how SVM works. It draws a decision boundary, i.e. a hyperplane between any two classes in order to separate them or classify them.

Now I know you're thinking how do you know where to draw a hyperplane?

The basic principle behind SVM is to draw a hyperplane that best separates the 2 classes. In our case the two classes are the rabbits and the wolves. Before we move any further, let's try to understand what a support vector is.

What is a Support Vector in SVM?

So, you start of by drawing a random hyperplane and then you check the distance between the hyperplane and the closest data points from each class. These closest data points to the hyperplane are known as support vectors. And that's where the name comes from, support vector machine.



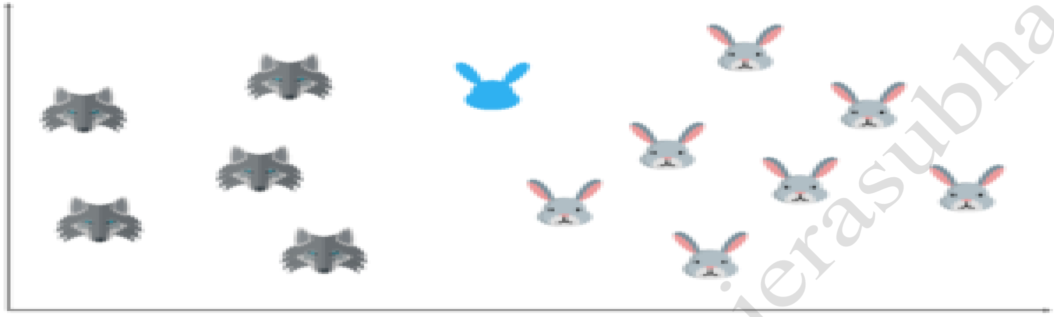
How does SVM work? – Support Vector Machine

The hyperplane is drawn based on these support vectors and an optimum hyperplane will have a maximum distance from each of the support vectors. And this distance between the hyperplane and the support vectors is known as the margin.

To sum it up, SVM is used to classify data by using a hyperplane, such that the distance between the hyperplane and the support vectors is maximum.

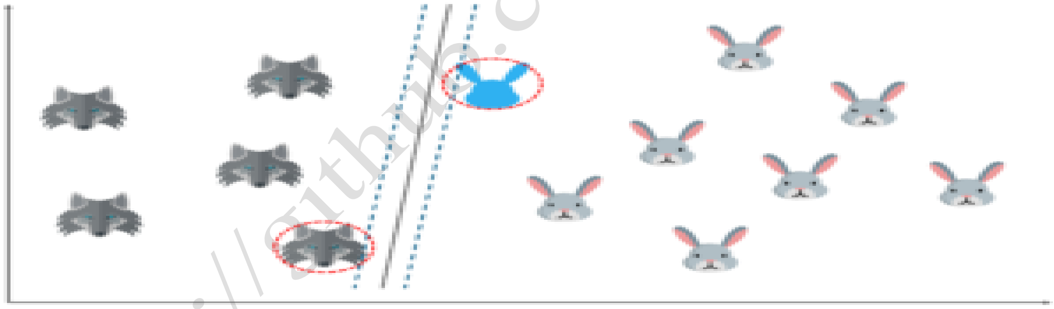
Alright, now let’s try to solve a problem.

Let’s say that I input a new data point and now I want to draw a hyperplane such that it best separates these two classes.



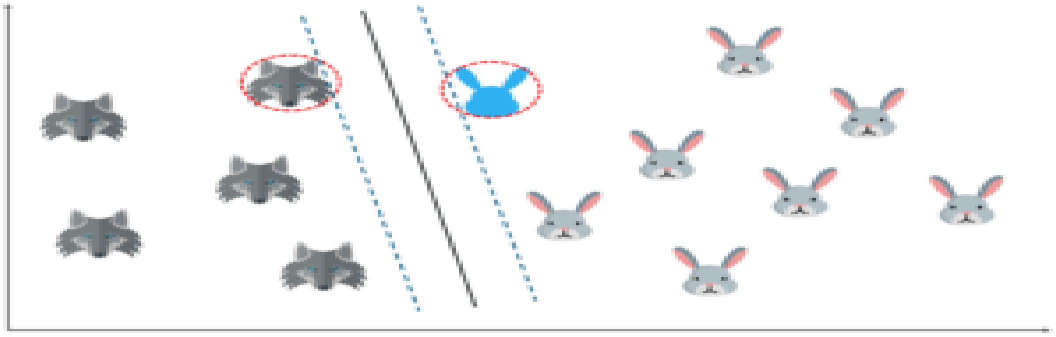
How does SVM work? – Support Vector Machine

So, I start off by drawing a hyperplane and then I check the distance between the hyperplane and the support vectors. Here I’m basically trying to check if the margin is maximum for this hyperplane.



How does SVM work? – Support Vector Machine

But what if I draw the hyperplane like this? The margin for this hyperplane is clearly more than the previous one. So, this is my optimal hyperplane.



How does SVM work? – Support Vector Machine

So far it was quite easy, our data was linearly separable, which means that you could a draw a straight line to separate the two classes!

1.6 Organization of Thesis

This dissertation organized into six chapters and structured as follows:

- *Chapter 1:*This chapter presents, “Introduction” of the thesis describing the Motivation, Problem Statement, Aim and finally organization of the dissertation.
- *Chaper 2:*This chapter presents, “Literature Survey” of the thesis describing about Classification Techniques in Healthcare using Data Mining and Machine learning approaches. Existing systems for liver disease classification and Related Work is seen.
- *Chapter 3:* This chapter presents, “System Architecture” of the thesis is describing the proposed System Architecture and System Modules.
- *Chapter 4:*This chapter presents, “Implementation” of the thesis describing about Dataset Collection, Data Analysis, Data Preprocessing, Data Visualization, Data Splitting, Model Building, Feature selection process,
- *Chapter 5:* This chapter presents, “Experimental results” of the thesis and the description of the results.

CHAPTER 2 LITERATURE SURVEY

In this section, detailed overview of the different research techniques are given with their working procedure.

Shambel Kefelegn et al [1] Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey. In this paper, dataset is collected from the uci data repository. Data cannot be given to the algorithm for training for this purpose data preprocessing is required. After preprocessing of data features are selected for the partitioning of the data in to training and testing data. Data mining technique for classification of patients with the liver disease and the patients who are not suffering from liver disease is applied such as SVM and NB. C4.5 Decision Tree considered for performance evaluation of liver disease classification using uci dataset. In this paper, classification experiments are performed using the Weka tool on the dataset. By comparing data mining classification algorithm, we got to know SVM accuracy score is 94.04% which is greater than NB using data mining approaches and 97.13% in machine learning which is also greater than NB approach. In this paper we got to know that SVM have the high performance rate when compared to NB while classification done using different domains such as machine learning and data mining. [1]

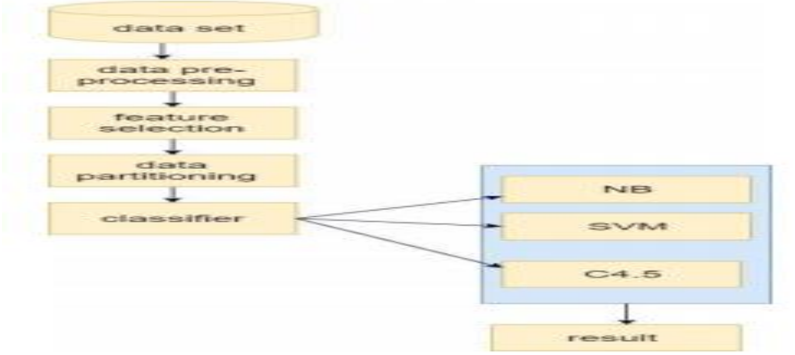


Fig2.1: Methodology Work Flow Diagram

K. Nagaraj and Amulyashree S et al[2] Neuro SVM: A Graphical User Interface for Identification of Liver Patients. In this paper, Indian liver patient dataset is collected from uci data repository. After the collection of data from uci data repository dual feature selection process is carried out. The features which are not in used at the time of training are removes which is know as correlated features. In this paper, based on the applications of the different algorithm feature are selected by using boruta package for classification are used such as NB, Bagging, Random forest and SMV . It is implemented using R platform. All the above algorithm are used for classification of the patients with liver disease. Classification is performed on ILPD dataset using the above algorithms. The classification here classifies the patients with the liver disease and patients who are not suffering from liver disease. Based on the performance rate of all the algorithms used in this paper, we got to know that SVM algorithm has the highest performance rate for classification when compared to all the above algorithm so in this paper hybrid neuroSVM model is developed and is constructed as a GUI using R platform.[2] The overall workflow of this paper is shown below in the figure.2.2.

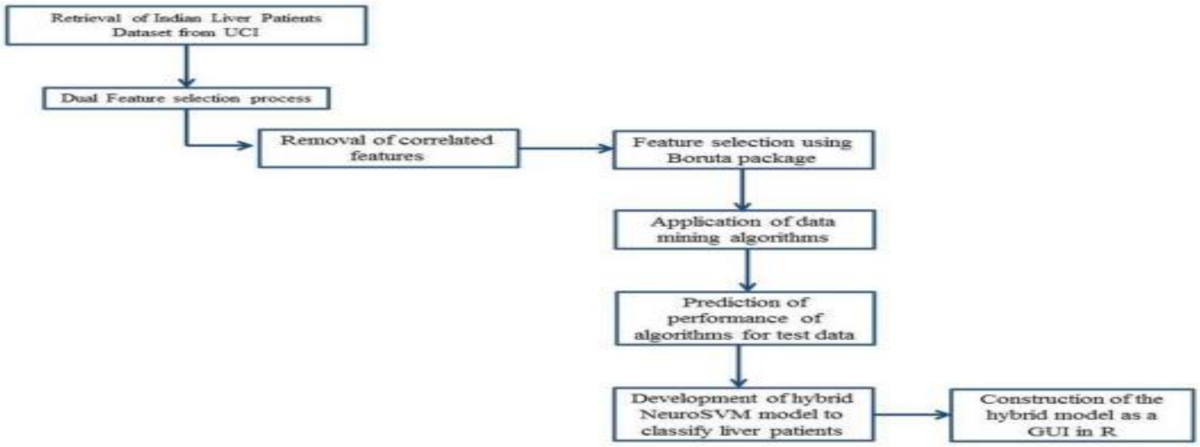

Retrieval of Indian Liver Patients Dataset from UCI
Dual Feature selection process
Removal of correlated features
Feature selection using Boruta package
Application of data mining algorithms
Prediction of performance of algorithms for test data
Development of hybrid NeuroSVM model to classify liver patients
Construction of the hybrid model as a GUI in R

Fig2.2: Flowchart for development of NeuroSVM model.

Harsha Pakhale and Deepak K et al[3] A Survey on Diagnosis of Liver Disease Classification. In this paper, ILPD(Indian liver Patient disease) is collected for training the model. As we cannot directly provide our data to our algorithm so preprocessing is required and integration of the dataset is done to improve the performance of our model.the next step is feature selection and feature transformation is done for training our model. Data mining techniques have been used for classification such as Decision tree, SVM, NB and Artificial Neural Networks for liver patient dataset. In the evaluation and presentation phase SVM and Artificial Neural rate has a highest performance rate in their patterns when compared to NB, Decision tress. Classification is done of patients with liver disease and no diagnosis is done.[3]

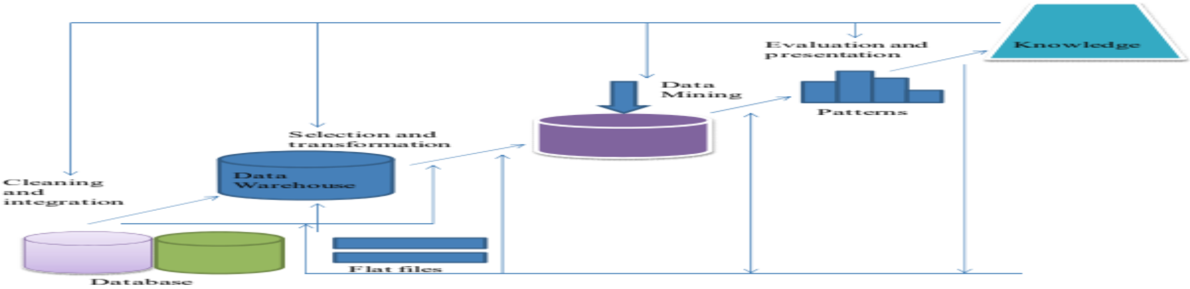


Fig2.3: KDD Process

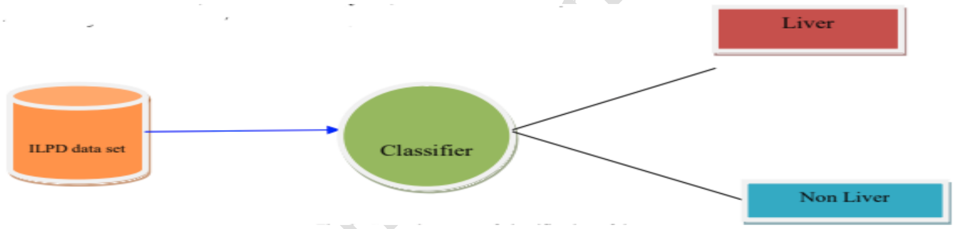


Fig.2.4: General process of classification of data

Kiran Kumar , M. Sreedevi and Y. C. A. Padmanabha Reddy et al[4] Survey on machine learning algorithms for liver disease diagnosis and prediction. In this paper, problem is defined for which the solution is shown in this paper. Data identification is done which is known as feature selection. Dataset splitting is done into two files such as training and tuning data. Training is used for training this model and tuning is used for testing the model. Based on features been selected machine learning algorithm such as supervised learning algorithms such as K-Nearest Neighbour and Support Vector Machine is used for classification of the liver disease patients. In the evaluation phase we got to know the performance rate of SVM is when compared to KNN.[4]

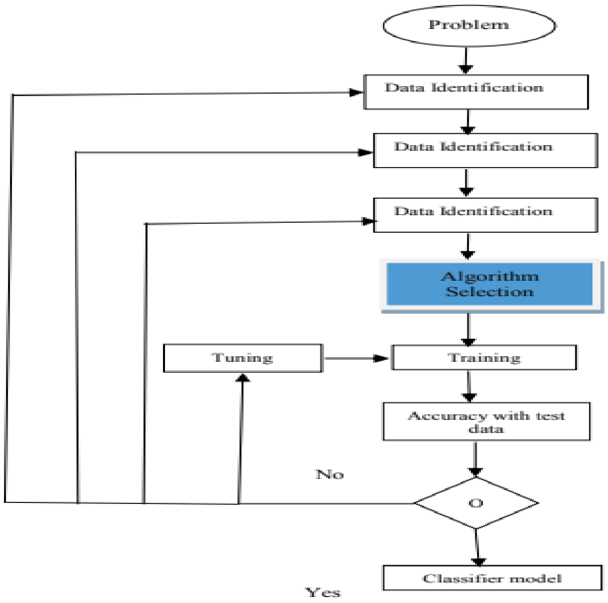


Fig.2.5: Supervised Learning Process.

Shakuntala Jatav and Vivek Sharma et al [5] An algorithm for classification data mining approach in medical diagnosis. In this paper, first phase of this model is preprocessing of the dataset. Data capturing, storing, Analyzing are the step of feature extraction. Data capturing is the collection of data which are necessary for our model. Storing data means features of the data which are important for model to train to analyze it. Based on the analysis Features are extracted from the dataset to train the dataset of this model. This research paper is mainly focused to classification of patients with liver disease possibility using data mining and machine learning approach in order to enhance the accuracy or precision of the disease detection expert system. This paper also shows the related work study of different approaches such as neural network, naïve bayes, SVM, KNN, FCN, etc and it is concluded that SVM gives the best performance as compared to the other existing techniques. These paper has designed using SVM and RF algorithms. The algorithms results with accuracy of 99.35%, 99.37 and 99.14 on diabetes, kidney and liver disease respectively. SVM have the high experimental results when compared to RF in this work.[5]

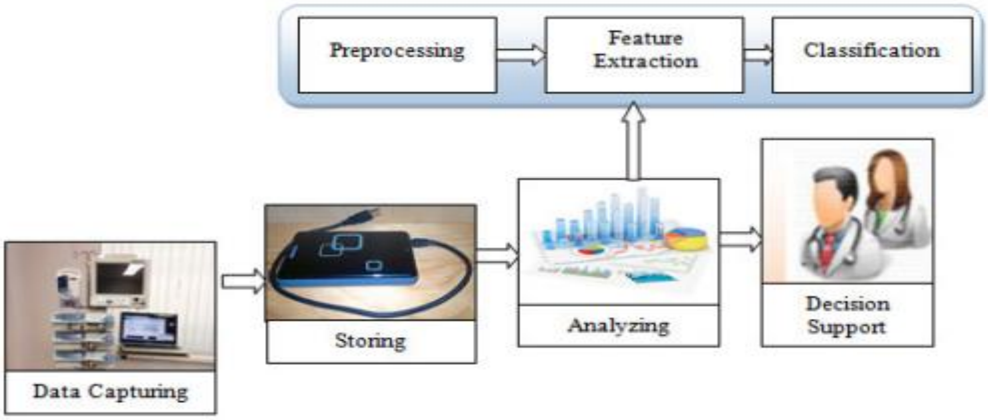


Fig2.6: A Typical Health Informatics Processing

K.Swapna and Prof. M.S. Prasad Babu et al [6] Critical Analysis of Indian Liver Patients Dataset using ANOVA Method. In this paper, ILPD data set is used for are Analysis of Variance (ANOVA). Analysis of different attributes such as gender, age and different liver function test such as SGOT, SGPT and Alkaline Phosphates. Each and every instance variable are analyzed in this project. Null values cannot be rejected in this method. Each value of every instance of the dataset is analyzed and studied carefully.[6]

Esraa M. Hashem and Mai S. Mabrouk et al [7]. A Study of SVM Algorithm for Liver Disease Diagnosis. Here in this paper, two datasets are collected from two different data repository such as uci and kaggle. After collection of the datasets SVM algorithm is used for classification of patients with liver disease using MATLAB software, in the evaluation phase performance of the two different dataset are evaluated. Accuracy, Error rate, sensitivity, Prevalence and Specificity shows the performance. Based on the evaluation phase we got to know the ilpd dataset attributes has the high performance when compared to bupa dataset attributes. This paper is the comparison of performance of two datasets such as BUPA dataset and ILPD dataset for classification using the svm algorithm. After classification evaluation of the classification is done based on confusion matrix. we get to know precision, recall, error rate and sensevity which describes the performance rate of both the datasets, in which we got to know that ILPD dataset performance rate is high. [7]

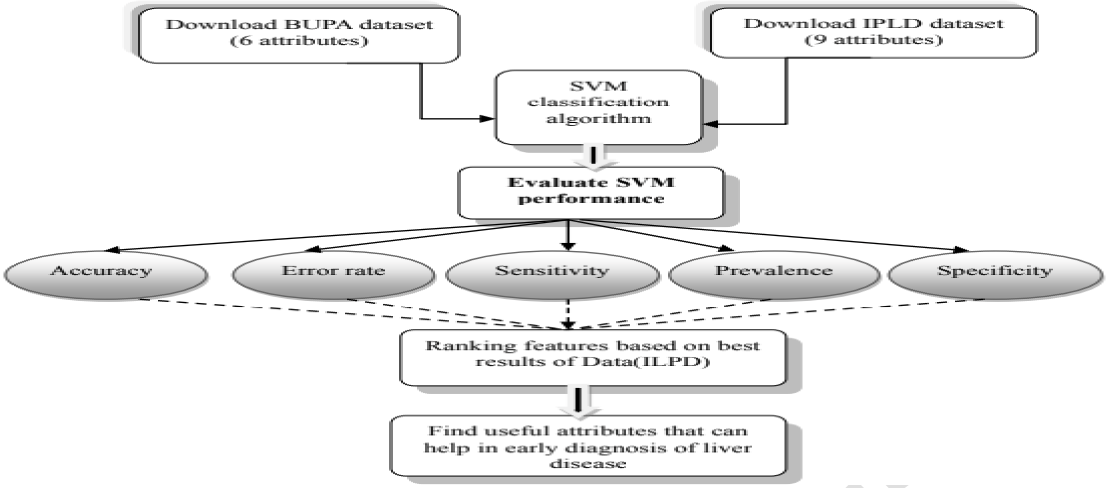

Download BUPA dataset
(6 attributes)
Download IPLD dataset
(9 attributes)
SVM classification algorithm
Evaluate SVM performance
Accuracy
Error rate
Sensitivity
Prevalence
Specificity
Ranking features based on best results of Data(ILPD)
Find useful attributes that can help in early diagnosis of liver disease

Fig2.7 The overall process

Nancy.P, Sudha.V and Akiladevi.R et al [8] Analysis of feature Selection and Classification algorithms on Hepatitis Data. This research paper mainly aims comparison of classification techniques on predicting the survival of the patients was done using feature selection algorithms. Hcc dataset is collected from the kaggle website. Data preprocessing is the essential step in every data mining project in order to provide the correct type of data to train the algorithm. Feature selection classification is carried out in the next phase. Evaluation of the performance of the different algorithms is done based on confusion matrix. Analysis provide a discovering of new patterns of disease in many medical datasets of disease occurrence and unearth the effect of data mining techniques in the medical arena. [8]

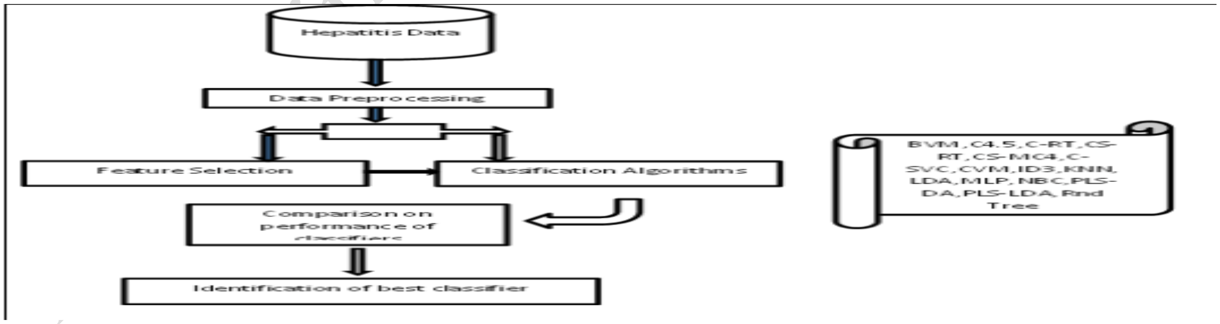

Hepatitis Data
Data Preprocessing
Feature Selection
Classification Algorithms
Comparison on performance of classifiers
Identification of best classifier
BVM ,C4.5,C-RT,CS-RT,CS-MC4,C-SVC,CVM,ID3,KNN, LDA,MLP, NBC,PLS-DA,PLS-LDA, Rnd Tree

Fig2.8 Framework of Data Mining Process

Dr. S. Vijayarani and Mr.S.Dhayanand et al [9] Liver Disease Prediction using SVM and Naïve Bayes Algorithms. In this paper dataset is collected from uci repository. Data mining techniques such as Classification is used here such as Naïve bayes and Support Vector Machine (SVM) for classification of patients with the liver disease and the patients who are not suffering from liver disease are classified. performance accuracy is checked of both the algorithms. The experimental results of this work concludes, the highest accuracy is observed in SVM when compared to Naïve Bayes classifier based on the execution time. [9]

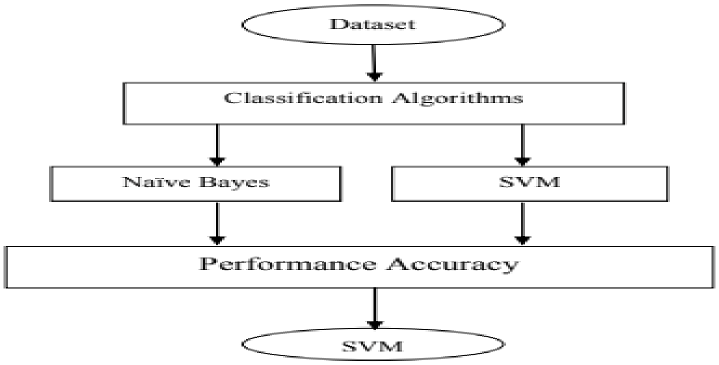


Fig2.9: System Architecture

Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew and Elizabeth Issac et al [10]. Diagnosis of Liver Disease Using Machine Learning Techniques. In this paper, they are using machine learning techniques such as SVM, Logistic Regression, KNN and Artificial Neural Network. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. ANN was the model that resulted in the highest accuracy with an accuracy of 98%. Comparing this work with the previous research works, it was discovered that ANN proved highly efficient. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using ANN. [10]

M. Banu Priya, P. Laura Juliet and P.R. Tamilselvi et al[11]. Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms. In this paper, structured and unstructured patient data is collected including all symptoms like fever,pain,pain in abdomen etc, neural network is applied, PSO(particle swarm optimization algorithm) feature selection methods is used for Indian Liver Patient Dataset. Decision support system and final diagnosis is done. The algorithms used in this paper are J48, MLP, SVM, Random Forest, and Bayes network Classification. Particle swarm optimization algorithm (PSO) various result for feature selection. It is observed that bayes network and J48 Classification gives better results compare to other classification algorithms. After which user interactive screen is developed. [11]

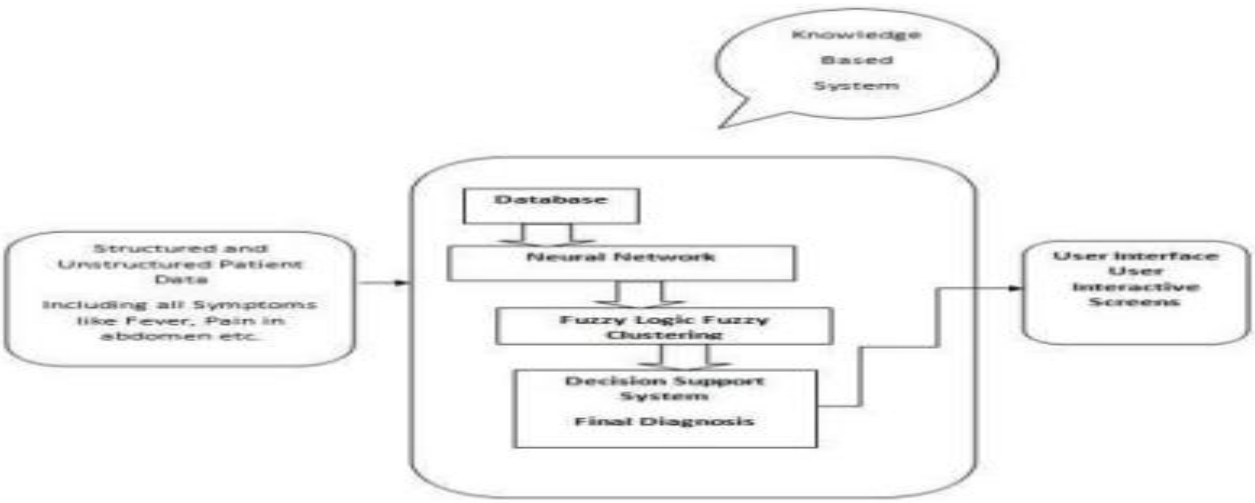

Knowledge
Based
System
Database
Structured and Unstructured Patient Data
Including all Symptoms like Fever, Pain in abdomen etc.
Neural Network
Fuzzy Logic Fuzzy Clustering
Decision Support System
Final Diagnosis
User Interface
User Interactive Screens

Fig2.10: Overall Architecture of Proposed System

The better classification can be guaranteed by comparing the different machine learning and data mining algorithm on different dataset. It is resolved by various researchers by introducing the different machine learning and data mining algorithm such as SVM, NB,C4.5 Decision Tree, Random Forest, J48, MLP and Bayesian Network are discussed. This work provides the detailed overview about the working procedure of multiple research techniques along with the merits and demerits. The overall evaluation of the research work is performed by comparing the working procedure and merits each method with other in terms of some performance metrics. This research work concluded with the better research method which can be applied to extract the useful information with the concern of context. Based on the performance rate SVM has high accuracy and is suitable classification of the patients with liver than the existing research methods.

CHAPTER 3 SYSTEM ARCHITECTURE

3.1 Proposed System

The proposed system is totally different from existing system, it consists of five different datasets for classification using two different algorithms. All the five dataset are based on the liver diseases. In two different datasets among the five the patients with and without liver disease are classified and in the remaining three datasets patient who will survive and who will die are classified. One among the two different datasets which are classified as patients with and without liver disease are further classified with the types of liver diseases the patients are suffering from.

There is no such automated tool or software used yet for the prediction of the liver diseases and till now liver diseases are predicted manually such as by the doctors and the pathology lab members. This system will surely be helpful across the world.

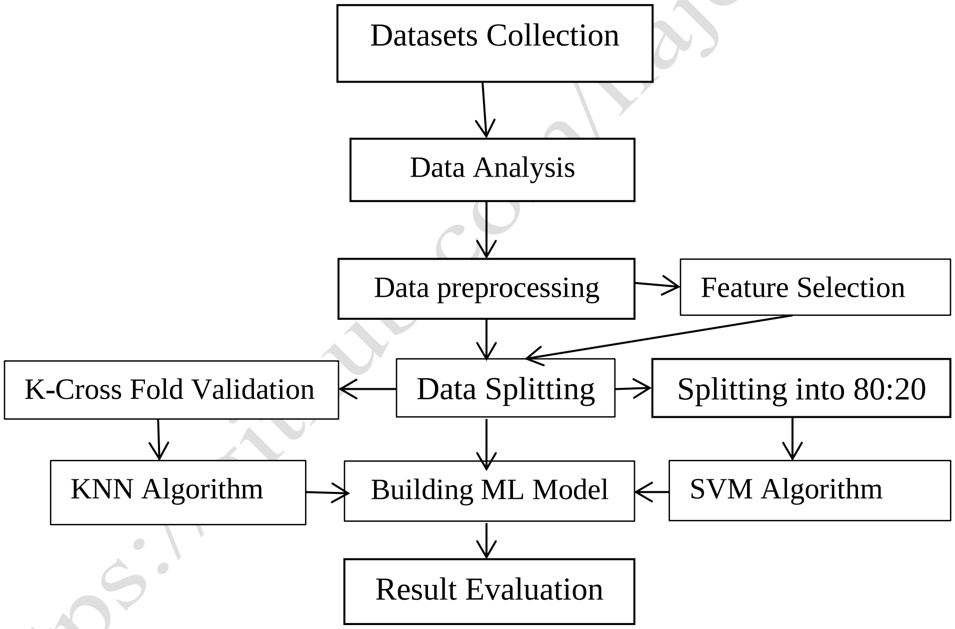


Fig: System Architecture

Module 1: Datasets Collection

Dataset 1: ILPD (Indian Liver Patient Dataset) Data Set: (Downloaded from kaggle) This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not) . This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

Since 1990, the JPAC Center for Health Diagnosis and Control, has conducted nationwide surveys of Indian adults. Using trained personnel, the center had collected a wide variety of demographic and health information using direct interviews, examinations, and blood samples. The data setconsists of selected information from 8,785 adults 20 years of age or older taken from the 2008–2009 and 2014–2015 surveys.

S.no	**Attribute**		**Attribute Description**			
1	Age	Numeric	Age of the patient			
2	Gender	Nominal	Gender of the patient			
3	TB	Numeric	Quantity of Total- Bilirubin			
4	DB	Numeric	Quantity of Direct Bilirubin in patient			
5	Alkphos	Numeric				
6	Sgpt	Numeric	Serum glutamic pyruvic transaminase (or) Alamine Aminotransferase in patient			
7	Sgot	Numeric	Serum glutamic oxaloacetic transaminase (or) Aspartate Aminotransferase in patient			
8	TP	Numeric	Total Proteins content in patient			
9	Albmin	Numeric	Amount of Albumin in patient			
10	A/G Ratio	Numeric	Albumin and Globulin Ratio in patient			
11	Class	Numeric	Status of liver disease in patient			

3.1 Proposed System

3.2 Architecture

CHAPTER 4 IMPLEMENTATION

Implementation is the very important phase of every machine learning project because in the implementation phase the decision or plan of the process is taken into effect; execution.

1.1 Dataset Collection

Data collection is the process of gathering and measuring data, information or any variables of interest in a standardized and established manner that enables the collector to answer or test hypothesis and evaluate outcomes of the particular collection.[techopedia]

4.2 Selection of Framework

4.3 Data Preprocessing

4.4 Data Analysis

4.5 Data Visualization

4.5.1 Barplot of age

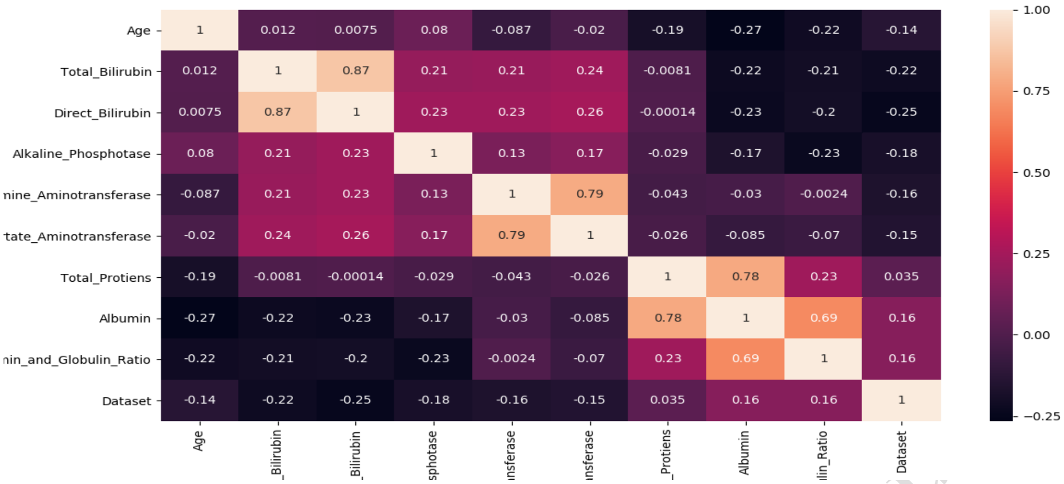
#importing the required dataset file

# Plot the bar chart for numeric values

# a comparison will be shown between

# plot between 2 attributes

4.5.3 Heatmap



4.6 Data splitting

4.7 Data Training Process

4.8 Data Modeling

4.9 Class Prediction

4.10 Prediction of different Liver Diseases

4.11 Deployment and Maintenance

REFERENCES

1. Shambel Kefelegn."Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey". International Journal of Pure and Applied Mathematics Volume 118 No. 9 2018.

2. Kalyan Nagaraj and Amulyashree Sridhar. "Neuro SVM: A Graphical User Interface for Identification of Liver Patients". PES Institute of Technology.

3. Harsha Pakhale and Deepak Kumar Xaxa. "A Survey on Diagnosis of Liver Disease Classification". MATS University,Raipur, Chhattisgarh, India . International Journal of Engineering and Techniques - Volume 2 Issue 3, May – June 2016

4. M. Kiran Kumar , M. Sreedevi and Y. C. A. Padmanabha Reddy. "Survey on machine learning algorithms for liver disease diagnosis and prediction". International Journal of Engineering & Technology, 7 (1.8) (2018) 99-102.

5. Shakuntala Jatav and Vivek Sharma. "An algorithm for predictive data mining approach in medical diagnosis". International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018

6. K.Swapna, Prof. M.S. and Prasad Babu et al. "Critical Analysis of Indian Liver Patients Dataset using ANOVA Method"
. International Journal of Engineering & Technology IJET-IJENS Vol:17 No:03 (2017)

7. Esraa M. Hashem and Mai S. Mabrouk. "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis". Biomedical Engineering, Misr University for

Science and Technology (MUST University), 6th of October 2014, Egypt
.American Journal of Intelligent Systems 2014.

8. Nancy.P, Sudha.V and Akiladevi.R. “Analysis of feature Selection and Classification algorithms on Hepatitis Data”. ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 1, January 2017.

9. Dr. S. Vijayarani and Mr.S.Dhayanand. “Liver Disease Prediction using SVM and Naïve Bayes Algorithms”. International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 4, April 2015.

10. Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew and Elizabeth Issac. “Diagnosis of Liver Disease Using Machine Learning Techniques”. International Research Journal of Engineering and Technology (IRJET)Volume:05,Issue:04, Apr-2018.

11. M. Banu Priya, P. Laura Juliet and P.R. Tamilselvi. “Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms”. International Research Journal of Engineering and Technology (IRJET)Volume:05,Issue:01 ,Jan- 2018

12. Miss. Hemalata Vaidya, Miss. S. K. Chaudhari and Mr. H. T. Ingale. “Literature review on liver disease classification”. Vol-3 Issue-3 2017.

13. Lawrence s. friedman(MD) ,Emmet b. keeffe(MD) *Stanford University Medical Center , Stanford, California* and Jules
L. dienstag(MD). “Handbook of Liver Disease”.

14. “Liver Function Testing in primary care” Developed by bpac,website: www.bpac.org.nz ,July 2007George St Dunedin, New Zealand

15. Christina Vett-Joice , “Capital Pathology” ,Published–2012.

16.B.R. Thapa and Anuj Walia, “Liver Function Tests and their Interpretation”, Indian Journal of Pediatrics, Volume 74— July, 2007

17. Machine Learning Techniques on Liver Disease - A Survey V.V. Ramalingam1 , A.Pandian2 , R. Ragavendran3 1,2,3Department of Computer Science and Engineering, SRMIST, Kattankulathur. International Journal of Engineering & Technology, 7 (4.19) (2018) 485-495 International Journal of Engineering & Technology Website: www.sciencepubco.com/index.php/IJET Research paper. Copyright © 2018 Authors.

18. https://www.yashodahospitals.com/liver-diseases/

19. https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/

20. https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/

21. https://towardsdatascience.com/ machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

22. https://www.edureka.co/

23.https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14

24.https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjL7pLWhv_iAhXZUn0KHbKpD5AQjRx6BAgBEAQ&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FNon-alcoholic-liver-disease-NAFLD-spectrum-NAFLD-encompasses-a-spectrum-of-fatty-liver_fig1_309962458&psig=AOvVaw1ybwyswSP9BTfpU-DHKe0s&ust=1561359700196924