# A Unified Machine Learning Framework for Effective Prediction of liver disease – Fatty Liver Towards Cirrhosis

By

**Attique ur rehman**

00000318850

Supervisor

**Dr. Wasi Haider Butt**

Department of Computer and Software Engineering

College of Electrical and Mechanical Engineering

National University of Science and Technology (NUST)

Islamabad, Pakistan

February 2022

# A Unified Machine Learning Framework for Effective Prediction of liver disease – Fatty Liver Towards Cirrhosis

By

**Attique ur rehman**

00000318850

Supervisor

**Dr. Wasi Haider Butt**

A thesis submitted in conformity with the requirements for the degree of *Master of Science* in Software Engineering

Department of Computer and Software Engineering

College of Electrical and Mechanical Engineering

National University of Sciences and Technology (NUST)

Islamabad, Pakistan

February 2022

# Declaration

I, *Attique ur rehman* declare that this thesis titled "A Unified Machine Learning Framework for Effective Prediction of liver disease – Fatty Liver Towards Cirrhosis" and the work presented in it are my own and has been generated by me as s result of my own original research.

Attique ur rehman,

00000318850

# Plagiarism Report

This thesis has been checked for Plagiarism. Turnitin report endorsed by Supervisor is attached.

Attique ur rehman

00000318850

Signature of Supervisor

# Copyright Notice

# Abstract

Liver is the largest organ of the human body with more than 500 vital functions. In recent decade number of liver patient has been reported such as cirrhosis, fibrosis, or other liver disorder. There is a need of effective, early, and accurate identification of individual suffering with such disease, so that the person may recover before the disease spread and become a fatal. For this, applications of Machine Learning are playing a significant role. In this research there are two main sub activities has been performed, firstly a protocol-based literature review (SLR) has been done. Secondly based on the SLR a unified Machine learning framework named as **Ma**chine **L**earning Based **Li**ver **D**isease **D**iagnose (MaLLiDD) has been formed.

In **SLR** phase we have reviewed 44 articles extracted from 5 different electronic repositories published from January 2015 to November 2021. After a systematic and protocol-based review we answered 6 research questions about machine learning algorithms. The identification of effective feature selection technique, data imbalance management technique, accurate machine learning algorithms, list of available data sets with their URL's and characteristics, and feature importance based on usage has been identified for diagnosing the liver disease. The reason to selecting this research question is, in any machine learning framework the role of dimensionality reduction, data imbalance management, Machine learning Algorithm with its accuracy and data itself is very significant.

MaLLiDD is a unified framework that has been tested on 3 different datasets to diagnose fatty liver towards cirrhosis and its severity level. The main contribution of this research is the framework MaLLiDD (that has been formed based on results of 44 studies), that can work for three different datasets helps in diagnosing liver disease. Performance given by MaLLiDD on cirrhosis diagnoses data set is 99.8 % , on ILPD Indian liver patient dataset 76.5% accuracy has been obtained and 76.1 % of accuracy on cirrhosis staging (severity level ) dataset.

**Keywords:** Classification, Machine Learning, Liver Disease, Cirrhosis, MaLLiDD, Unified Framework, Fatty Liver, SLR.

# Acknowledgment

All praise to Allah (the omnipotent and the omnipresent) who has bestowed me with ardor, courage, and patience with which I have completed another phase of my academic journey.

I would like to dedicate this thesis to *my beloved younger brother* who passed on December 17, 2021, after a prolong fight (3 and half years) against cancer. I would also dedicate this thesis to my parents, teachers, siblings, and students, who were continuous source of motivation in my tough times. They always encouraged me to continue higher studies and fully supported me to full fill my dream degree. My parents played a pivotal role in my MS degree by providing moral and financial support.

Most importantly, I want to pay special gratitude to my supervisor “Dr. Wasi Haider Butt” without whom I would not have been able to take this task to fruition. He enlightens my path with continuous support and made me competent during the whole duration of research.

Finally, I am thankful to “Dr. Waseem Anwar”, “Dr. Sabeen Javaid”, “Dr. Tahir Muhammad Ali” and all my friends who assisted me in this thesis and throughout the whole research process.

# Table of Content

## List of Figures

# List of Tables

# CHAPTER 1

# INTRODUCTION

The functionality of liver is strongly affected by viral diseases that cause inflammation. Such disease can get severe and shown as a fatal one. Liver Disease has been observed as a common clinical disorder that is being increased in parallel with diabetes, alcohol, obesity, and metabolic syndrome.[1] Liver disease prevalence in higher number is also a burden on economy. There is a need of early and accurate identification of humans with higher risk so treatment plan can be made as early as possible.[2]

According to World Health Organization in men the liver cancer is the fifth most common cancer and 9th in the women. In a single year of 2018 840,000 cases of liver cancer were reported while the number was 782,000 in 2012. The mentioned diagnoses were happened to the patients aged above 75 years. The prognosis of liver cancer is very poor due to mortality to incidence is 0.95% [3]. Liver consists of two main lobes that may divided into further eight functionally independent segments. Liver cancer are mainly of two types of primary and secondary. The Cancer begins in cells of the liver called primary liver cancer, these cancerous cells join to form a lump or start its growth at multiple sites. Its growth happened exponentially. Hepatocellular (HCC / hepatoma) and cholangiocarcinoma (bile duct cancer) are categorized under primary liver cancer.[4]

Hepatitis C virus (HCV) and Hepatitis B virus (HBV) are one of the common causes of cirrhosis. Patients are mostly asymptomatic in early cirrhosis as liver has the ability to compensate[5]. But this compensation cannot go for long and its exponential growth can be fatal in near time. There is a need of quick and early diagnosis with treatment of this chronic disease, delay in diagnosis and treatment can eventually leads to complication and mortality[6] [7].

There has been a significant advancement in Machine Learning technology result in application of Machine learning in various aspects of life. Medical Field is being highly benefitted by application of Machine Learning especially in diagnosis process.[8] Machine learning helps in diagnosis of liver disease using multiple parameters. The outcomes of algorithms help doctors and lab practitioners to assess the presence of liver disease before it becomes fatal [9]. The process of screening the liver patient has become easy with the help of machine learning. There are many classification algorithms that can classify the normal persons and persons with chronic liver disease[10]. Machine learning algorithms also refers as data-mining algorithms or tool, that aims to recognize patterns existed in dataset. The algorithms make sure to classify the relevant instance in their classes or groups.[11] The more instances classified in a correct class the more accurate algorithms are. While using machine learning in medical field there is a high risk as health is being involved in the whole process, and accuracy of algorithm is a thing to focus on. For example, if machine learning algorithms declare a liver patient as a normal one, it'll be more dangerous for patient [12]. There are many states of art machine learning algorithm available. We cannot prioritize anyone except based on accuracy shown on a certain dataset. In some cases, one algorithm can be best while same algorithm may be worst for other dataset. Some common algorithms available namely K-nearest neighbor, Naïve Bayes, Decision Tree, Random Forest, Deep learning, Support Vector Machine, Logistic Regression [13].

## 1.1. MOTIVATION

The liver organ is located at upper right portion of abdominal cavity, top of the stomach, intestine and right kidney. The shape of liver is like a cone, colored dark reddish brown having weight of 3 pounds as shown in Figure 1: Structure of Liver. Liver gets supply of bloods through two distinct ways 1) oxygenated blood flows in from the hepatic artery 2) nutrient-rich blood flows in from the hepatic portal vein. [14]In amount liver holds a 13 percent of the total blood of any human body. There are two main lobes in liver having 8 segments each and each segment with 1,00 lobules.[15]

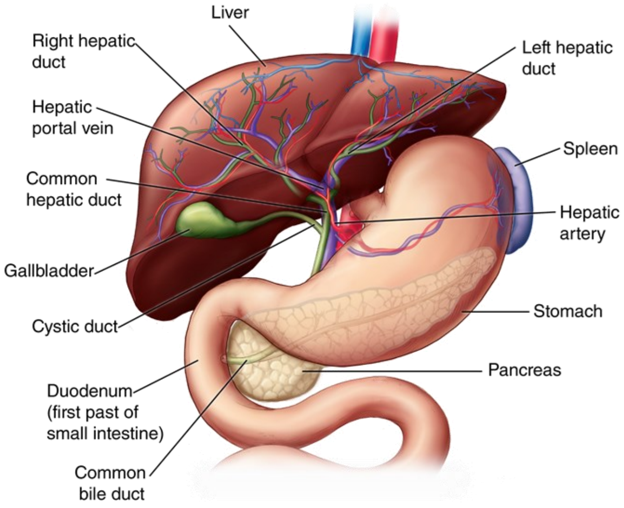


*Figure 1: Structure of Liver*

Liver aims to regulate or balance most of the present chemicals in human body and convert it into bile. In this way liver carry away the waste product from itself. The blood leaving the stomach and intestines passes through the liver. The blood further processed to classify the nutrients that are nontoxic and easier to use for the whole body. The chemical / waste that is not useful for body got away from the blood to be exit from body. More than 500 main roles of liver organ have been identified. [16]

- ✓ Bile production to carry away waste
- ✓ Proteins production for blood plasma
- ✓ Cholesterol production
- ✓ Conversion of Glucose into glycogen and back to glucose again when required
- ✓ Regulating levels of amino acids
- ✓ Hemoglobin processing

- ✓ Poisonous ammonia conversion to urea
- ✓ Cleansing blood from drugs and poisonous substance
- ✓ Regulation of blood clotting
- ✓ Removal of bacteria from blood
- ✓ Removal of bilirubin (excess bilirubin turns skin and eyes to yellow)

Whenever the harmful substance broken down by the liver its remaining passed to either as bile or blood. Bile further passes through intestine and exit from body whereas blood filtered by the kidney and exits from the body.

## 1.2. PROBLEM STATEMENT

It is really difficult to diagnose the chronic Liver Disease, and classify it into the respective class, it is cirrhosis, fibrosis etc. moreover the severity level prediction is also difficult job manually. There is a need of a technique that can predict liver disorder, check whether it is cirrhosis with its severity level. With the mentioned approach the probability of a patient survival may increase. Medical experts like doctors, laboratory practitioners and researchers can be benefitted from this research.

## 1.3. AIMS AND OBJECTIVES

Main objectives of this research activity are mentioned categorically in this section

### *1.3.1. Systematic Literature Review Objectives*

- ✓ To identify Machine Learning algorithms that are being used for diagnosis of liver disease.
- ✓ Accuracy of Machine Learning algorithms diagnosing liver disease.
- ✓ Identification of data sets being used to diagnose liver disease.
- ✓ Effective feature selection techniques being used to diagnose liver disease.
- ✓ Identification of data imbalance management techniques.
- ✓ List of most important features.

### *1.3.2. Framework Objectives*

- ✓ To help doctors in diagnosing patient with liver disorder
- ✓ To classify patients as liver cirrhosis positive or negative
- ✓ To classify the cirrhosis patient as severe or less severe
- ✓ A unified framework that can do all three above mentioned jobs
- ✓ To do above mentioned job in highest accuracy specially diagnosing cirrhosis patients

## 1.4. STRUCTURE OF THESIS

The rest structure of thesis is followed as:

Chapter 2

Chapter 3

Chapter 4

Chapter 5

CHAPTER 2

# MACHINE LEARNING AND LIVER DISEASE

## 2.1. MACHINE LEARNING OVERVIEW

Machine learning is the one of evolving branches of computational algorithms, aims to emulate intelligence of human by learning from the real-world environment. The Algorithms refers as working horse in modern era of BIG DATA. There are vast applications of machine learning algorithm in big data. The data can be related to any field ranging from computer vision, finance, signal processing, computational biology, entertainment, spacecraft engineering, defense technology, chemical analysis, and others. 50 percent of cancer patients receives radiotherapy that requires a prescribed dosage of radiotherapy. Such process involves human machine interaction and machines are trained based on previous patterns. There are some tasks that machines can do in a perfect manner as compared to human. The machine learning algorithm is a process of computation which uses a data set as input (based on which some tasks are assigned as outcome) to algorithm

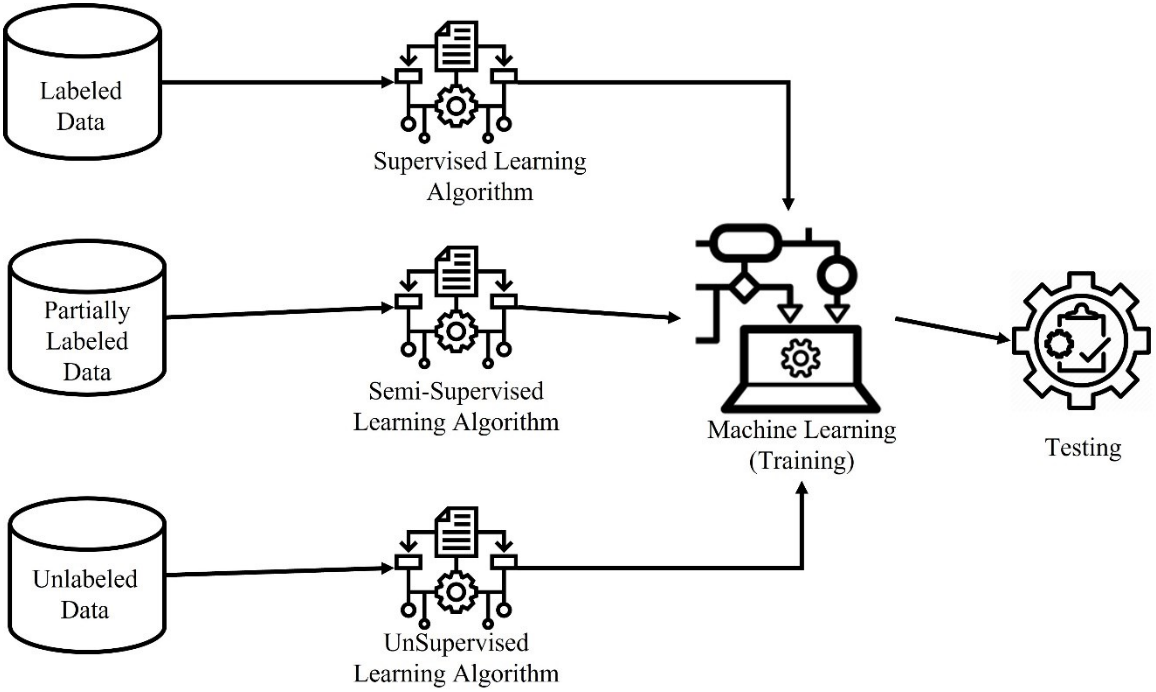

Labeled Data
Supervised Learning Algorithm
Partially Labeled Data
Semi-Supervised Learning Algorithm
Machine Learning (Training)
Testing
Unlabeled Data
UnSupervised Learning Algorithm

*Figure 2: Machine Learning Overview*

without being hard coded. Through repetition they can adapt their architecture based on learning, that’s why they are called soft coded. The algorithms repeat their instructions number of time unless they produced the required outcome.

Machine learning is referring as a technology that is used to develop computer-based algorithms that can emulate human intelligence. Machine learning is a branch of Artificial Intelligence but can be drawn from various other domains like statistics, control theory, philosophy, psychology, and information theory. There are three main types of machine learning algorithms. *Supervised, Unsupervised, Semi-Supervised* machine learning algorithms. In supervised learning, labelled data is used as input for training once the supervised learning-based model trained, then unlabeled data given as input to test output based on which algorithm can be checked how accurate it is. Example of classifying bananas and oranges. In unsupervised learning data given as input is totally unlabeled, algorithm must made clusters of data instances based on their features and dimensions. The most relevant instances will be clustered together without any label. Number of clusters always given by the practitioner whereas algorithm produced required number of clusters. In Semi supervised algorithm partially labeled data given to the algorithm. Means there are some data instances that are labeled, and others are unlabeled in this scenario labeled data helps to make unlabeled one as a labeled one as shown in Figure 2: Machine Learning Overview.

## 2.2. MACHINE LEARNING IN BIOMEDICINE

The idea of assistance by computation in improving the field of medicine is old as a digital computer. The scientists have already used computer in diagnosing blood disease during 1960s. The results derived using machine learning algorithms may used by researchers, medical officers like doctors to study the disease pattern in different and efficient way. It also helps in making advancement of medical treatment based on diagnostic results. This is not obvious always that a data scientist does the job alone there may be a need of medical officer to collaborate with data scientist so the statistical analysis on data may perform in a better way. Machine learning can help in number of ways to medical officers. Some of which are

- Reconstructing Disease
- Hypothesis testing
- Recruiting patients
- Developing diagnostics
- Improving prognostics
- Patient monitoring

### *2.2.1. Reconstructing Disease*

The machine learning helps in reconstruction of certain disease pattern with the help of unlimited computing power. The clinical officers use the bulk of existing data for such analysis in details they apply certain drugs to read the reaction of drug dose on that disease. For each disease there is a certain pattern that needs to be broken to cope up with disease, the practitioners with the help of simulations tried to break that pattern. For example, in 2019 a virus named as COVID-19 hit badly the whole world, during that period scientist used the computing powers to break the covid-19 disease patterns resulted in production of vaccination that is being used worldwide after testing. Again, the initial testing may get done on computers using simulators with big data enabled high power computing.

### *2.2.2. Hypothesis Testing*

In any medical field initial structure or model to predict the outcome is quite easy without testing unless it is tested. While testing it becomes the most difficult job. One without any biological background knowledge can make a statistical model using machine learning to predict or diagnose the disease. Once the statistical analysis-based model prepared the closed collaboration required between data analytic engineer and medical officer, because medical officer better knows which dimension can be ignored and which to be used in any case.

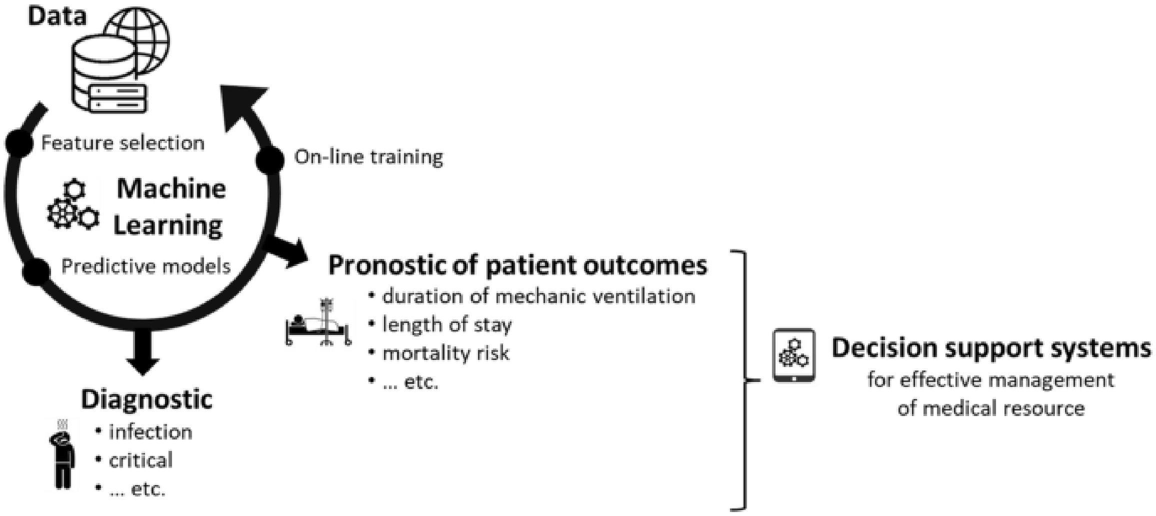


*Figure 3: A Machine Learning Application in Biomedicine*

### *2.2.3. Recruiting patient*

It is the most difficult activity to recruit patients for the certain trials. Most of the time the trial ends without any results due to absence of patient sample. Machine learning helps in recommending possible sample patient for a particular trial with the help of previous medical records. In Figure 3: A Machine Learning Application in Biomedicine has been shown that is used for decision support system.

## 2.3. LIVER OVERVIEW AND DISEASES

Any condition that can affect or damage liver's normal functionality is called liver disease. It can be classified into many severities level normally in 4 classes ranging from normal to severe one. Over time if liver left untreated it can lead to cirrhosis.

### 2.3.1. *Liver Organ Overview*

The only largest organ of human body is liver. The major function of liver is to remove toxins and harmful chemicals from body and regulates the blood supply within body. It also maintains or regulate blood sugar levels and blood clotting. The liver organ is located at upper right portion of abdominal cavity, top of the stomach, intestine and right kidney. The shape of liver is like a cone, colored dark reddish brown having weight of 3 pounds the exact location of liver is shown in Figure 4: Location of Liver in Human Body. Filtration of all blood in body is primary responsibility of liver organ. The poisonous substances like drugs and alcohol filtered out from the blood and exit from the body with the help of liver. The liver produces bile, which is fluid like, helps in digesting the food and make waste exit from the body. The primary aims of liver is to produce Bile, Albumin, filter blood and regulate the blood clotting, resist infections, storing vitamins and minerals, process glucose and regulation of amino acids. These are the key functions there are many other functions being done by liver organ.

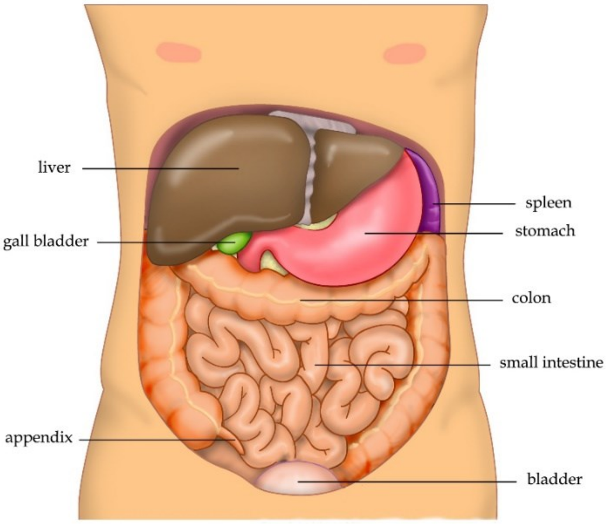
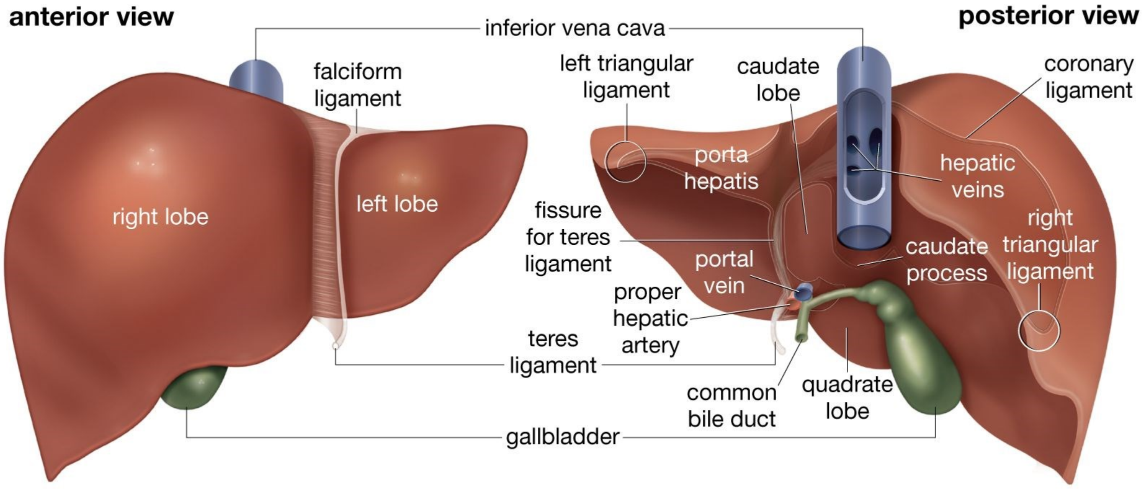


*Figure 4: Location of Liver in Human Body*

### 2.3.2. *Structure of the liver*

The liver organ consists of four lobes the right bigger in size lobe, the left lobe, and the smaller quadrate and caudate lobes. As shown in Figure 5: Labelled Liver Structure by Encyclopedia Britannica, Inc, The right and left lobe is divided by falciform ligament. The falciform ligament further connects the liver to abdominal wall. The lobes of liver further divided in to eight smaller segments that are made up of thousands of smaller lobules. Each lobule (smaller segment) has a duct flowing towards the common hepatic duct that helps in draining bile from the liver. The parts in liver are named as Common Hepatic Duct to carry bile out of the liver, Falciform ligament to separate two lobes of liver, Glisson's Capsule a loose connective tissue surrounds liver, Hepatic Artery the main blood vessel through which liver gets oxygenated blood, Hepatic Portal Vein to carry blood to liver, Lobes the anatomical sections of liver, Lobules a smaller microscopic block building of liver, Peritoneum membrane covering the liver.



© 2010 Encyclopædia Britannica, Inc.

*Figure 5: Labelled Liver Structure by Encyclopedia Britannica, Inc*

## 2.3.3. Liver Diseases

Approximately 2 million deaths per year occurred due to liver disease worldwide from which 1 million only due to liver cirrhosis and other due to hepatocellular and hepatitis. There are many kinds of liver disorder or diseases from which some are enlisted below.

- ✓ Liver disease caused by viruses like hepatitis A, B and C.
- ✓ The liver disorder caused by poisons, drugs or too much consumption of alcohol. For example, cirrhosis and fatty liver disease.
- ✓ Disease that is inherited genetically including Wilson Disease, Hemochromatosis.

## 2.3.4. Liver Cirrhosis

The condition in which liver is scarred and permanently damaged is known as liver cirrhosis. These scar tissue replaces the healthy one and eventually the whole liver organ replaced by such scarred tissue that prevents liver from its normal functionality. If such disease left untreated sooner the disease become fatal and takes individual's life. 11th most common cause of death in world is Liver cirrhosis while liver cancer is 16th leading cause of death globally. On combining liver cancer and cirrhosis are death cause of 3.5 % of all deaths globally. In world there are around 2 billion humans consuming alcohol from which 75 million are diagnosed as liver related disorders. In common organ transplantation liver organ transplantation is on 2nd position [5]. This is not only burden on economy of any country but also a big question mark on public health. Patient initially face fatigue and severe itchy skin. As discussed earlier, liver has an ability of compensation, if not diagnosed earlier the whole liver may badly damage. The common causes include high consumption of alcohol, nonalcoholic fatty liver, hepatitis A, B and C. There are many diagnosis methods including liver function test (LFT), viral infections testing, imaging tests and liver biopsy. There is not any solid treatment that can reverse cirrhosis effect, only physician can do is to restrict its growth in liver and make sure liver does not failed. In Figure 6: Healthy Liver vs cirrhosis patient's Liver , the difference between normal liver and cirrhosis patient's liver has been shown. It is obvious and one can see the scarred and infected liver in worse condition.

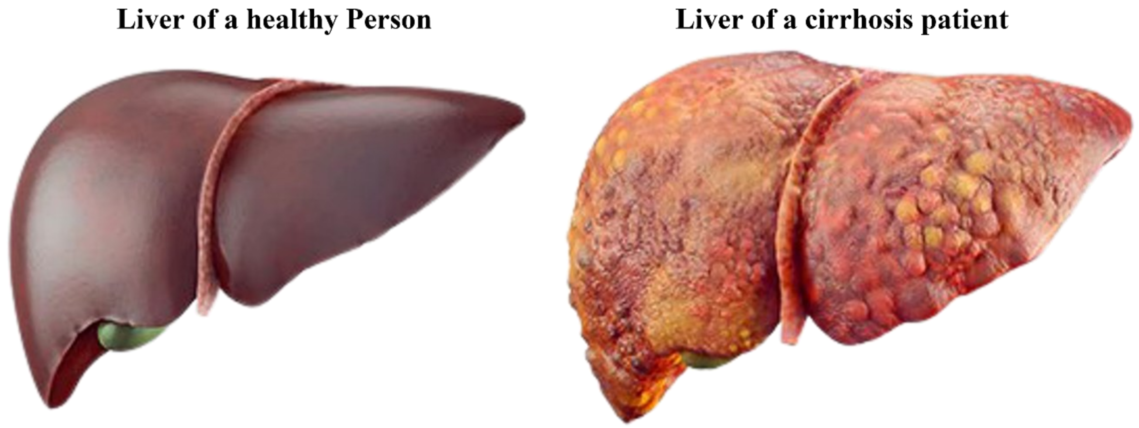


*Figure 6: Healthy Liver vs cirrhosis patient's Liver*

There are number of precautions that one should consider avoiding from being infected by such disease some of which enlisted below.

- Totally avoid consumption of alcohol or consume in a certain limit.
- Food or drinks that contain high fructose corn syrup or trans fats should be discouraged and avoid.
- Carefully intake the medical prescription such as drugs or other injections that directly hits liver causing liver injury.
- Exercising on regular basis may reduce risks
- Consume as less as you can, the red meat.

One should immediately contact to a doctor if color of urine or stool got changed, jaundice or yellowing of eyes, pain in the upper right of abdomen and on swelling of arms or legs.

### Chapter Summary

In this chapter, application of machine learning in domain of biomedicine has been discussed with background knowledge of liver disease advancing towards the cirrhosis. Application of machine learning that helps medical society has been discussed in detail. Major functions of liver with its structure and anatomy have been discussed.

CHAPTER 3

# SYSTEMATIC LITERATURE REVIEW

SLR is a systematic way to explore the existing literature for a certain problem. The said approach is based on some certain phases already discussed in introduction section. Using SLR approach makes your work unbiased, and one can rely on the findings of the literature review. The SLR approach proposed by Kitchenham and Charters [17] is good enough that many researchers in domain of computer and software engineering adopted it and gave their significant contribution using this approach. The suggestions that had been provided by Kitchenham and Charters [17] are significant to get used in this research contribution. The proposed phases consist of planning, conducting, and reporting the literature review. Mentioned phases already discussed in Figure 7: Systematic Literature Review Phases.

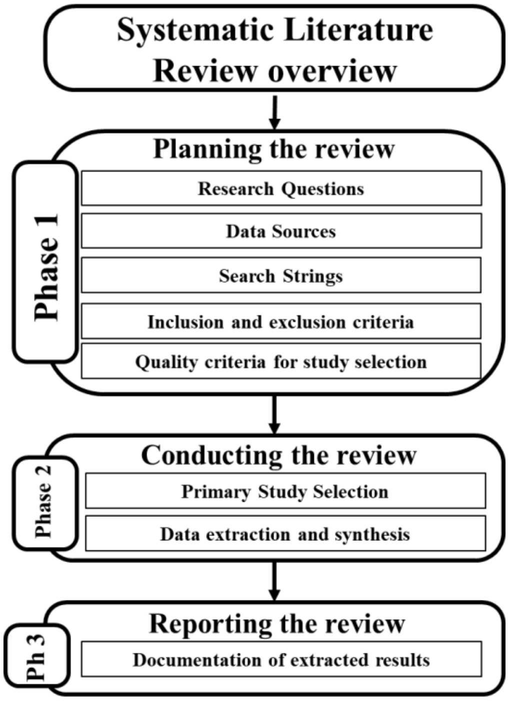

Systematic Literature Review overview
Phase 1
Planning the review
Research Questions
Data Sources
Search Strings
Inclusion and exclusion criteria
Quality criteria for study selection
Phase 2
Conducting the review
Primary Study Selection
Data extraction and synthesis
Ph 3
Reporting the review
Documentation of extracted results

*Figure 7: Systematic Literature Review Phases*

## 3.1. PLANNING REVIEW

Based on our objectives we proposed research questions to be answered. The answers will directly help us to propose our own solution that will better than existing solutions. To get answers to these questions we selected the related data repositories from where we can download or read related literature. After selection of data repositories search strings has been defined so that most related literature can be searched from electronic repositories. Then definition of inclusion and exclusion criteria has been done. Final articles will be scanned through quality assessment criteria, so that the high-quality articles may include in our final research, and the answers to research question may be in the highest quality.

### *3.1.1. Research Questions*

In this study main objective is to find existing datasets, their applications, machine learning algorithms than can be applied and validation techniques. In this context following research questions has been designed that are directly related to main objectives of current study.

**RQ1:** Which Machine Learning algorithms are being used for diagnosis of liver disease?

**RQ2:** How accurately Machine Learning algorithm diagnosing liver disease?

**RQ3:** What are the data used to diagnose liver disease?

**RQ4:** How feature selection is being done on data to diagnose liver disease?

**RQ5:** How imbalance data has been dealt. Which is the best method?

**RQ6:** What are the most important features?

### *3.1.2. Data sources*

Appropriate and highly related electronic data repositories has been identified. The mentioned repositories are related enough that the research objective may fulfilled. The electronic data repository is enlisted in Table 1: Electronic Repositories Data search sources

*Table 1: Electronic Repositories Data search sources*

| | |
|---|---|
| Electronic Data repositories | PubMed (https://pubmed.ncbi.nlm.nih.gov/).<br>Wiley Inter Science (https://onlinelibrary.wiley.com/).<br>ACM Digital Library (https://dl.acm.org/).<br>IET digital library (https://digital-library.theiet.org/).<br>Springer Link (https://link.springer.com/).<br>IEEE Xplore (https://ieeexplore.ieee.org/).<br>ScienceDirect (https://www.sciencedirect.com/). |
| Searched items | Full text articles including Journals, conferences, and workshops |
| Language | English |
| Publication period | From January 2015 to November 2021 |
| Searching process | Formulation of search strings to extract the papers that are highly related. |

### *3.1.3. Search strings*

Research questions helped to formulate the search strings. Specific keyword and their alternatives have been extracted from research questions. All keyword and alternatives have been formulated based on existing literature of Liver disease and its diagnosis. Final search string has been developed using logical AND, OR operators. Formulated search strings have been enlisted as Table 2: Used Search String. The final query will be very feasible to search the most relevant literature in current domain. The logic operators made the query most optimistic.

*Table 2: Used Search String*

| Keywords | Alternatives |
|---|---|
| Machine Learning (K1) | (“Machine Learning” OR “Deep Learning” OR “Artificial Intelligence”) |
| Diagnosis (K2) | (“Diagnosis” OR “Detection”) |
| Liver Disease (K3) | (“Liver Disease” OR “Fatty Liver” OR “Fibrosis” OR “Cirrhosis”) |
| Prediction (K4) | (“Prediction” OR “Supervised Learning” OR “Classification” OR “Grouping” OR “Ensemble”) |
| Dataset (K5) | (“Dataset” OR “Features” OR “Feature Selection” OR “Data of patients”) |

The Final formulated search string has been developed as: (K1) AND (K2) AND (K3) AND (K4) AND (K5)

### *3.1.4. Inclusion criteria*

An appropriate criterion has been defined before including research papers in current study. Complete inclusion criteria have been defined below in bullets.

✓ Paper to be included should be published as journal, conference, workshop, or a book chapter.
✓ The paper that discussed diagnosis of liver disease using any machine learning algorithm.
✓ The paper that discussed accuracy of diagnosis and results in details, so that comparison may made.
✓ Papers that have been published between (January 2015 to November 2021)
✓ Studies that have been published in English language.

### *3.1.5. Exclusion criteria*

Below defined the criteria to exclude irrelevant literature from extracted research papers.

✓ Paper that are pre-prints or not peer reviewed.
✓ Papers that are not contributing to current study objectives.
✓ Papers that do not discussing diagnosis of liver disease.
✓ Papers that do not using machine learning algorithms to diagnose liver disease.
✓ If duplication of any study found the most current and complete published version will be used and rest all will be discarded.

### *3.1.6. Study quality evaluation*

Final literature has been evaluated based on quality evaluation criteria after paper extraction phase. To evaluate selected articles a check list has been created as mentioned in Table 3: Quality Evaluation Criteria of Selected Studies. The study quality checklist consists of 5 study quality questions based on which the study quality score has been assigned. If any study answers all SQ questions (SQ1 – SQ5) the study quality score will be 5 however on partial answers 2.5 will be given and 0 upon no answer. The final selection will be made based on benchmark study quality score and will consider as the most contributing studies in term of objectives of SLR.

*Table 3: Quality Evaluation Criteria of Selected Studies.*

| Study Quality Score | Study Quality Score Criteria |
|---|---|
| SQS-1 | "The studies were given '5' score that answering all the questions mentioned in the checklist" |
| SQS-2 | "The studies were given '2.5' score that partially answering the questions mentioned in the checklist" |
| SQS-3 | "The studies were given '0' score that are not answering any of the questions mentioned in the checklist" |

| SQ Questions | Study Quality questions checklist |
|---|---|
| SQ1 | "Does the study is addressing the mentioned research questions" |
| SQ2 | "Does the study used any machine learning methods" |
| SQ3 | "Are the study is related to any liver disease" |
| SQ4 | "Does the study include validation of proposed solution" |
| SQ5 | "Does the study provide comparison of results to any other study" |

## 3.2. CONDUCTING THE REVIEW

Phases of conducting the review has been further divided in sub phases as follows.

### *3.2.1. Primary study selection*

The articles were further refined using tollgate approach proposed by Afzal W et. al [18]. The method tollgate approach consists of five major phases as shown in Table 4: Tollgate Approach phase wise table and in Figure 9: Tollgate approach for article selection.

*Table 4: Tollgate Approach phase wise table*

| EDR | ST-1 | ST-2 | ST-3 | ST-4 | ST-5 | % (N=44) |
|---|---|---|---|---|---|---|
| PubMed | 120 | 103 | 049 | 018 | 006 | 13.6 % |
| Wiley Inter Science | 719 | 595 | 393 | 246 | 002 | 04.5 % |
| ACM Digital Library | 143 | 104 | 054 | 38 | 007 | 15.9 % |
| Google Scholar | 775 | 644 | 127 | 46 | 010 | 22.7 % |
| IEEE Xplore | 122 | 097 | 070 | 24 | 019 | 43.2 % |
| Total | 1879 | 1543 | 693 | 372 | 44 | 100 % |

EDR = Electronic Data Repository, ST = steps, % percentage with respect to all studies

At start, 1879 articles were extracted from online repositories using search strings composed using Boolean operators between keywords and their alternates. The total articles were further refined based on phase wise inclusion and exclusion criteria. The final 44 (2.34 % of total) articles have been included based on tollgate approach for further statistical analysis and to answer the specified research questions. The refined articles were further checked for quality assessment based on quality criteria results shown in Appendix A reference.

### *3.2.2. Data extraction and synthesis*

To answer the research questions, data extraction process has been defined using kitchenhem approach [17]. For which tables has been designed using spreadsheet. The data extracted is following

- ✓ Study labels Appendix A reference
- ✓ Year of Publication
- ✓ Type of study journal conference
- ✓ Only machine learning methodology based extracted

All articles were carefully reviewed to check the methodology used, accuracy of methodology, model evaluation and other data mining features that must be mentioned in articles. The validity of study has been checked and addressed in study quality table reference.

## 3.3. REPORTING THE REVIEW

Based on extracted studies this section includes the reporting of review with results.

### *3.3.1. Quality attributes*

Study quality question were prepared to check the quality of study. The selected studies with their quality score have been listed in Appendix A. The threshold for inclusion of study in research was ≥ 80%. Study quality score has been granted based on how study participating in answering to the research questions. The more quality score means more related to our research goals and should be investigated further to target the research questions.

### *3.3.2. Temporal distribution of selected primary studies*

All selected papers have been categorically divided in two categories either it is a journal publication or a conference publication, after which articles has been divided based on the year of publication. It has been seen that out of 44 articles 24 belongs to conference publication and 20 from journals. Most of the papers published in year 2020 as shown in

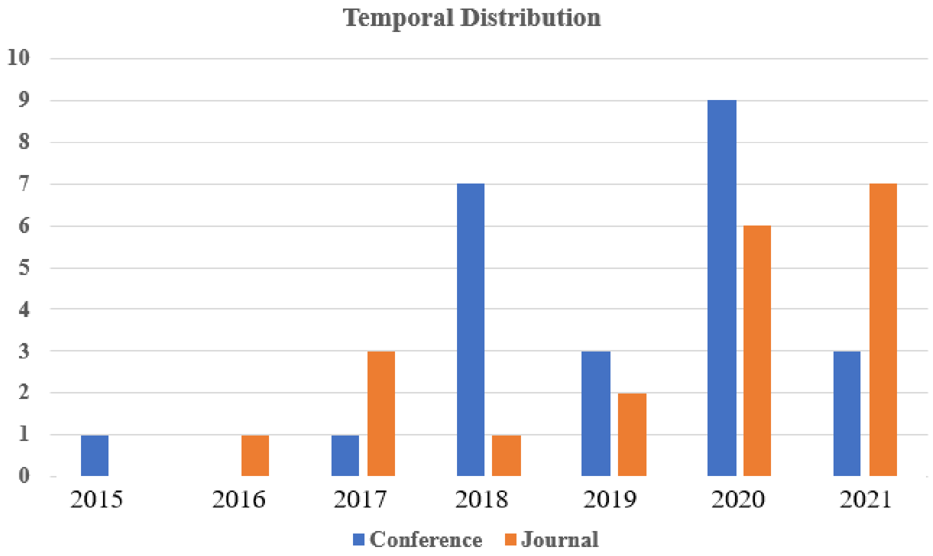


*Figure 8: Temporal distribution of selected articles*

Figure 8: Temporal distribution of selected articles and Table 5: Temporal Distribution with conference and journal count. Graph showing the illustration of temporal distribution based on category journal and conferences with their respective count.

*Table 5: Temporal Distribution with conference and journal count*

| | **Conference** | **Journal** | **Count** |
|---|---|---|---|
| **2015** | 1 | 0 | 1 |
| **2016** | 0 | 1 | 1 |
| **2017** | 1 | 3 | 4 |
| **2018** | 7 | 1 | 8 |
| **2019** | 3 | 2 | 5 |
| **2020** | 9 | 6 | 15 |
| **2021** | 3 | 7 | 10 |
| **Total** | **24** | **20** | **44** |

The graph has been illustrated using the table. in which brief details about count of years has been listed. Number of journal and conference papers in a year has also enlisted to analyze year and journal / conference count.

### *3.3.3. Used research method in selected studies*

The primary goal of this study was to answer the research questions about machine learning algorithm used to diagnose the liver diseases with their characteristics. So only the paper with machine learning methodology has been extracted, also discussed in inclusion and exclusion criteria. Paper that used data mining / machine learning approach were refined further to answer the research questions. All 44 papers included in this study has used the machine learning approach / methodology to diagnose the liver disease.

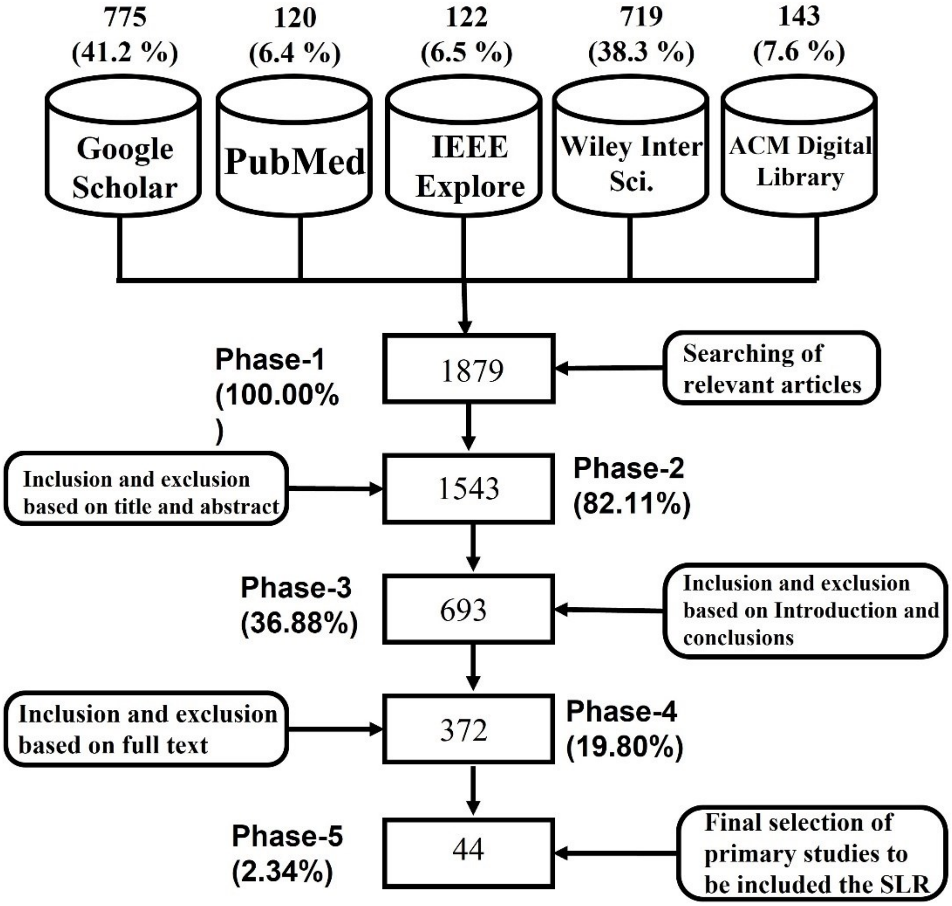

775
(41.2 %)
120
(6.4 %)
122
(6.5 %)
719
(38.3 %)
143
(7.6 %)
Google Scholar
PubMed
IEEE Explore
Wiley Inter Sci.
ACM Digital Library
Phase-1
(100.00%
)
1879
Searching of relevant articles
Inclusion and exclusion based on title and abstract
1543
Phase-2
(82.11%)
Phase-3
(36.88%)
693
Inclusion and exclusion based on Introduction and conclusions
Inclusion and exclusion based on full text
372
Phase-4
(19.80%)
Phase-5
(2.34%)
44
Final selection of primary studies to be included the SLR

*Figure 9: Tollgate approach for article selection*

## 3.4. RESULTS AND DISCUSSION OF SLR

Complete results relating to the research questions has been discussed in this section. All 44 articles were carefully analyzed, and results has been extracted and discussed in this section. The results are totally based on research questions.

### *3.4.1. Identified Machine Learning Algorithm*

The selected study has been carefully analyzed to extract most used machine learning algorithms. As shown in graph Random Forest has been used most of the time in selected 44 studies. 5,4,6,8,2,10 times Decision Tree, Logistic Regression, Artificial Neural Network, K-Nearest Neighbors, Support Vector Machine and Random Forest respectively.

Out of 44 studies 35 used 6 Machine Learning algorithms mentioned in table, rest 9 studies used different or special algorithms or changed the existing one. The most frequent algorithm has been extracted and enlisted as a table below. It has been observed that some researchers used ensemble of more than one algorithm such approach has been included in calculation in such a way that if ensemble KNN and Decision tree has been used then usage count of KNN and Decision Tree, both increased by one. There are some studies that apply more than one algorithm individually on datasets and then made comparison, the only algorithm that shows best result has been included in algorithm usage table in this study.

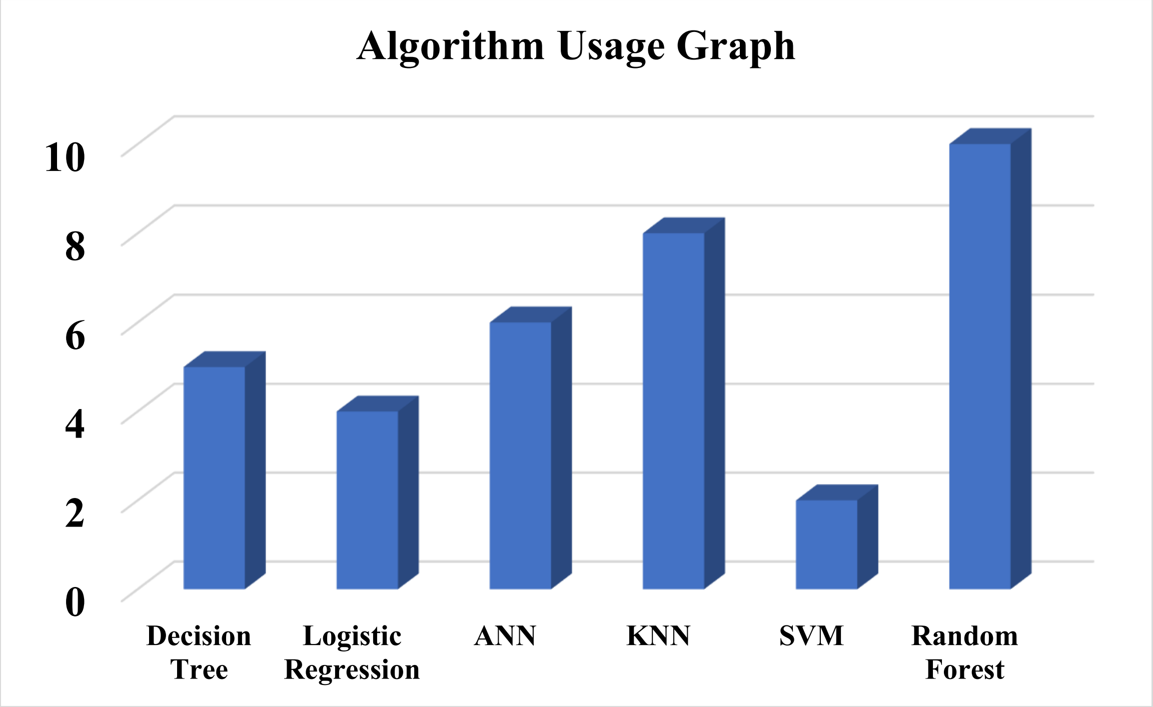


*Figure 10: Algorithm used for classification*

*Table 6: Algorithm Usage Count*

| Sr. | Algorithm | Usage |
|---|---|---|
| **1** | Decision Tree | 5 |
| **2** | Logistic Regression | 4 |
| **3** | Artificial Neural Network | 6 |
| **4** | K Nearest Neighbors | 8 |
| **5** | Support Vector Machine | 2 |
| **6** | Random Forest | 10 |
| | **Total** | **35** |

### *3.4.2. Accuracy achieved by Machine Learning Algorithms*

Based on accuracy achieved by all refined 44 studies average accuracy for each algorithm has been calculated so that a decision can be made that which algorithm performed better than other ones. After careful calculations average accuracy achieved by Decision Tree, Logistic Regression, Artificial Neural Network, K-Nearest Neighbors, Support Vector Machine, and Random Forest is 86.74%, 75.065%, 84.47%, 80.461%, 86.635% and 88.893%.

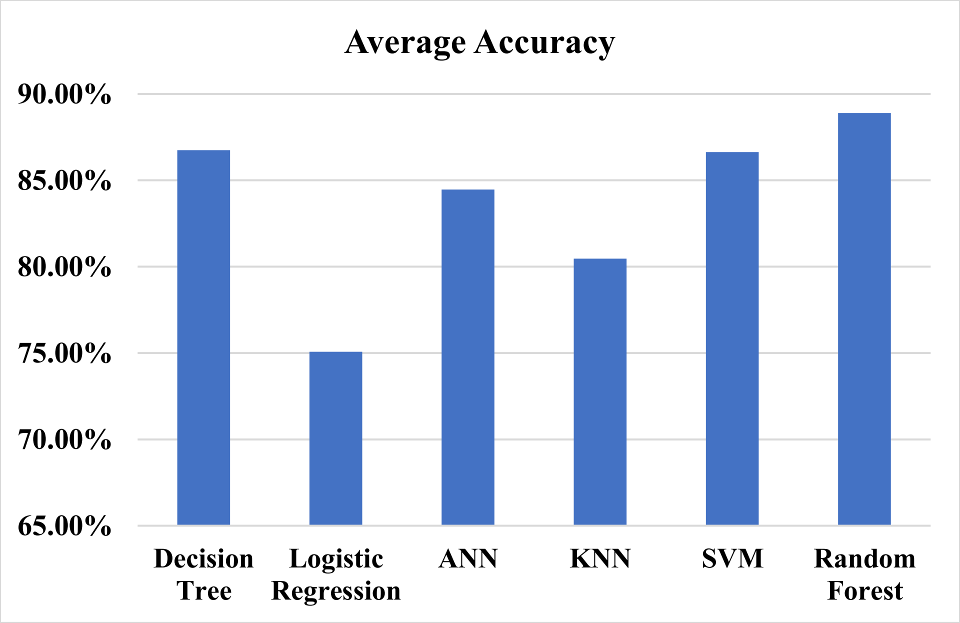


*Figure 11: Average accuracy achieved by each Algorithm*

Random forest shown best average accuracy, also it is considered as the most used machine learning algorithm according to this literature review, as it has been used by 10 authors to evaluate their machine learning models for classification. The method used to calculate average accuracy is 1) extract individual algorithm accuracy from papers 2) take sum of accuracy of a particular algorithm 3) divide on number of usages same as taking arithmetic mean.

*Table 7: Average accuracy by individual algorithm table*

| Sr. | Algorithm | Avg Accuracy |
|---|---|---|
| **1** | Decision Tree | 86.74% |
| **2** | Logistic Regression | 75.065% |
| **3** | Artificial Neural Network | 84.47% |
| **4** | K nearest neighbors | 80.461% |
| **5** | Support Vector Machine | 86.635% |

| | | |
|---|---|---|
| **6** | Random Forest | 88.893% |
| **Full Literature Average** | | **83.71%** |

After analyzing full literature average accuracy of all machine learning algorithm has been calculated as 83.71%. Based on usage it has been seen that the random forest is the most significant and its average accuracy is extracted from 10 readings, while the least significant machine learning algorithm is support vector machine as its average accuracy has been extracted from 2 readings only. To obtain accurate and meaningful results in terms of arithmetic mean the no of readings matters.

### *3.4.3. Identified important features*

To identify most important features from all datasets a detailed analysis has been performed, based on which below table has been designed. Total 22 features have been enlisted in table. It has been seen that same feature has been included under a different column name in different datasets. In table below this issue has been normalized and all alternative names has been writer in same cell. For example, feature 1 (F1) has been included in different datasets with name SGOT / AST / Aspartate Aminotransferase.

*Table 8: Table of Identified important features*

| **Features** | | [19] | [20] | [21] | [22] | [23] | [24] | [25] | [26] | [27] | [28] | [29] | [30] | [31] | [32] | [33] | [34] | **Usage** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **F1** | (SGOT) (AST) Aspartate Aminotransferase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 12 |
| **F2** | Total Bilirubin | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 12 |
| **F3** | (SGPT) (ALT) Alanine Aminotransferase | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 11 |
| **F4** | Alkaline Phosphatase | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 9 |
| **F5** | Albumin | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 8 |
| **F6** | Age | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 8 |
| **F7** | Direct Bilirubin | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| **F8** | Gama glutamyl transferase | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 5 |
| **F9** | Total Proteins | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **F10** | Gender | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 4 |
| **F11** | Body Mass Index | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 4 |
| **F12** | Albumin and Globulin Ratio | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| **F13** | Choline Esterase | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| **F14** | Cholesterol | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 3 |
| **F15** | Alpha-fetoprotein | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 3 |
| **F16** | creatinine | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| **F17** | TG Triglycerides | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 |
| **F18** | Hyaluronic Acid | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **F19** | Alpha 2-Macroglobulin | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **F20** | Apolipoprotein | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **F21** | Uric Acid | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| **F22** | High-Density Lipoproteins | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

22 features F1 to F22 has been extracted from selected studies. These feature names have been extracted either from the link of data sets or the names written in research papers. Some papers don't mentioned name of features that has been used for results, so that studies are excluded from this analysis.

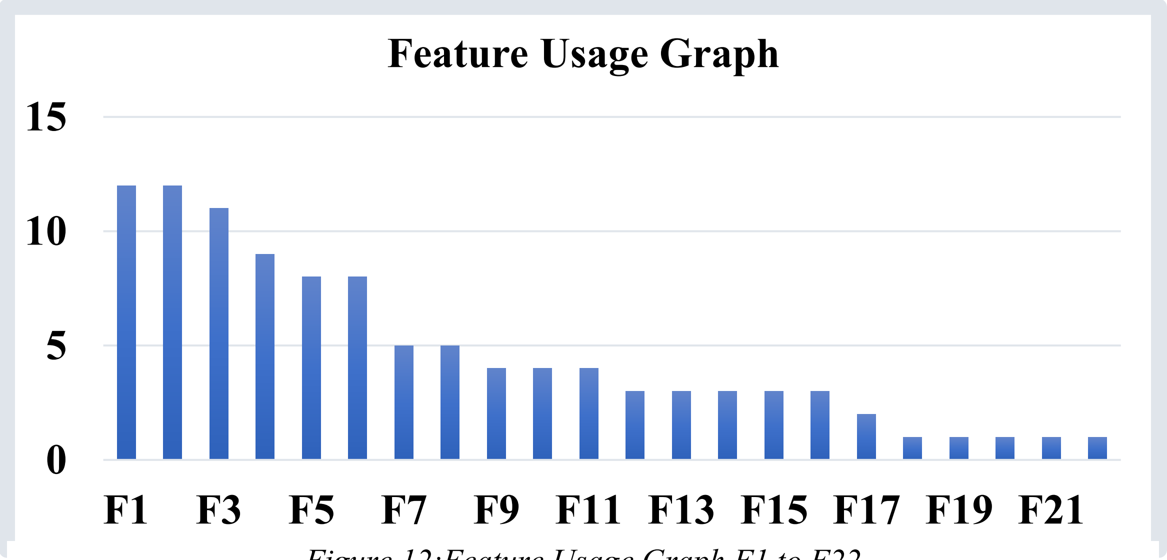


*Figure 12:Feature Usage Graph F1 to F22*

**Studies** [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34]

The mentioned studies used one of the F1 to F22 feature in their final features for results, so the analysis has been done based on these studies. To find important features, all features were enlisted mentioned by the extracted 16 studies. For each feature, all 16 studies have been analyzed, whether the study has included the feature in the feature set based on which results has been obtained. If the feature is in list of feature set that are used for results, then the count for that feature increased by one otherwise it remains zero. So, for every feature, every study has been checked and count has been placed in intersecting cell of study and feature.

A final usage column has been included in table to show how many studies included a particular feature. For example, feature 1 (F1) SGOT and F2 has been used by 12 studies for results. So, the F1 and F2 is the most significant and overlapped feature. One can set a specific threshold based on usage column and select the number of features to do an effective analysis. Based on specific threshold initially some features from dataset can be selected. It seems that the features F1 to F6 are the most significant and one cannot ignore these features in doing classification.

The above analysis of features can help in reducing time for feature selection. For example, if someone is using brute force for feature selection where data consists of more than 10 features. There will be minimum 1024 iterations to be done for optimized features to be selected. Thus, this is a minimum number this number can be increased exponentially due to increase in number of features as $2^n$ where n is the number of features to be go through from brute force cycle. This feature analysis shows that SGOT / AST / Aspartate Aminotransferase and Total Bilirubin is the most significant based on the refined 44 studies as these features has been used by 12 authors in their final feature selection. Rest features as SGPT / ALT / Alanine Aminotransferase used by 11, Alkaline Phosphatase by 9, Albumin and age by 8, Direct Bilirubin and Gama glutamyl transferase by 5, Total proteins, Gender and Body Mass Index by 4 only, Albumin and Globulin Ratio, Choline Esterase, Cholesterol, Alpha-fetoprotein, and creatinine has been used by 3,3 authors respectively. If we set a threshold of usage count to 8 then F1 to F6 are the most significant.

## *3.4.4. Identified data imbalance management techniques*

In various real-world application, the problem of imbalanced data is inevitable such as classification of images, recognition of patterns, signal processing, recognition of abnormal activity and diagnosis / prognosis of fatal diseases. As this study is targeting the liver patients, so a problem of imbalanced data may result in declaring a liver patient as a healthy one which can be fatal for patient. As we know that the early diagnosis may help to increase chances of recovery on treatment. One should deal imbalanced data before applying any machine learning algorithm for classification. For this number of data imbalance management techniques has been identified from literature and illustrated in Figure 13: Identified Data Imbalance Management Technique Usage illustration

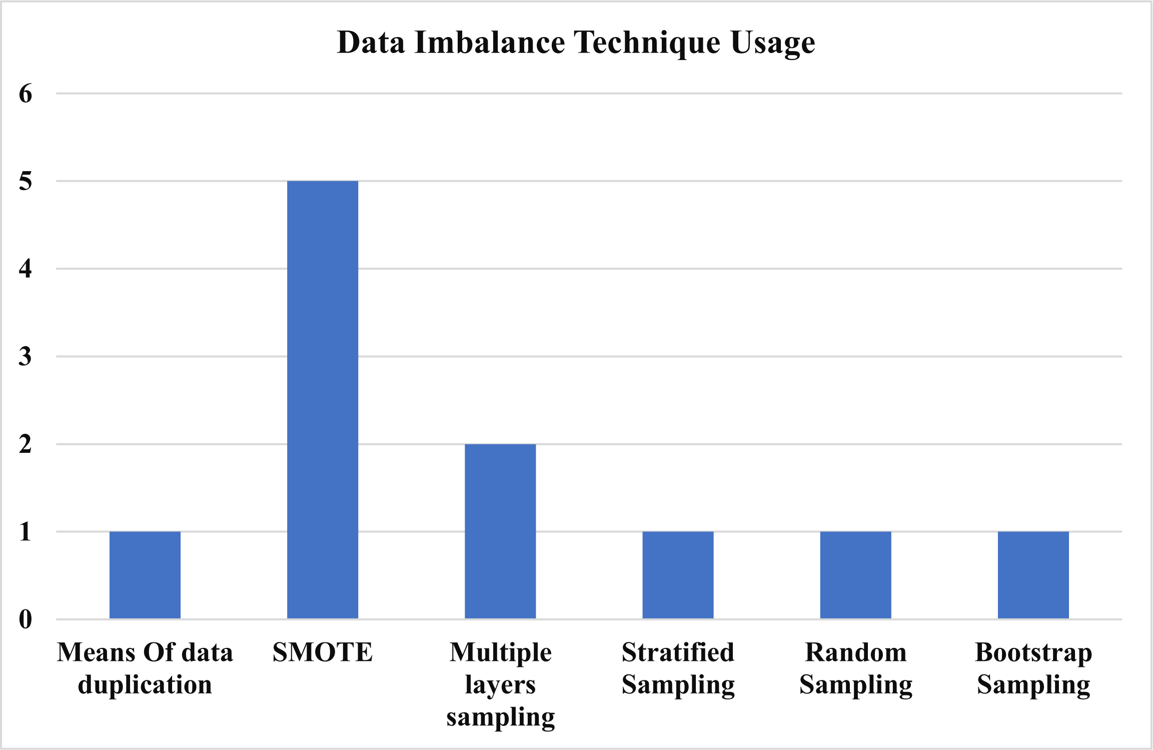


*Figure 13: Identified Data Imbalance Management Technique Usage illustration*

From 44 articles 6 data imbalance management techniques has been identified such as means of data duplication that has been used by 1 author, similarly, SMOTE by 5, Multiple layers sampling by 2, stratified sampling by 1, Random sampling by 1 and Bootstrap Sampling by 1 author. It has been observed that Synthetic Minority Oversampling Technique has been used by 5 authors which shows its significance. The Table 9: Table of Data Imbalance Technique with Usage shows usage count of each Data imbalance technique with its usage.

*Table 9: Table of Data Imbalance Technique with Usage*

| SR. | Data Imbalance | Usage |
|---|---|---|
| **1** | Means Of data duplication | 1 |
| **2** | SMOTE | 5 |
| **3** | Multiple layers sampling | 2 |
| **4** | Stratified Sampling | 1 |
| **5** | Random Sampling | 1 |
| **6** | Bootstrap Sampling | 1 |
| | **Total** | **11** |

### *3.4.5. Identified feature selection techniques*

In past decade, the large growth of datasets production has been observed. Whereas the application of machine learning needs to be highly accurate and process in learning the pattern existing in data. More number of dimensions cause more complexity in data. There is a need of feature reduction so that the highest accuracy and speed can be achieved. One of the most used feature reduction techniques is feature selection. This section of SLR study aims to extract most effective feature selection technique based on its usage on liver disease patient. The Figure 14: Identified Feature Selection Technique shows the number of extracted feature selection technique with its usage by different authors. It has been observed that Correlation Based feature selection has been used by most of the authors, count is 11 showing its significance. Total 9 feature selection techniques have been extracted. The details of feature selection technique have been enlisted in Table 10: List of feature selection technique with usage.

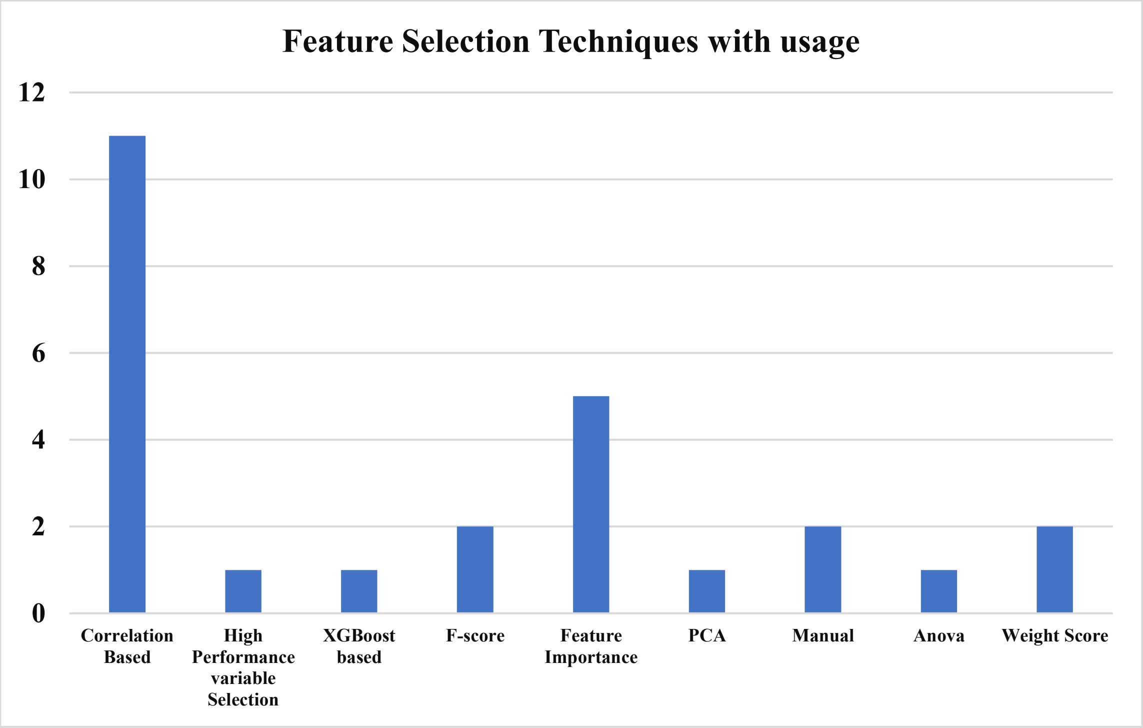


*Figure 14: Identified Feature Selection Technique*

*Table 10: List of feature selection technique with usage*

| SR. | Feature Selection | Usage |
|---|---|---|
| **1** | Correlation Based | 11 |
| **2** | High Performance variable Selection | 1 |
| **3** | XGBoost based | 1 |
| **4** | F-score | 2 |

| | | |
|---|---|---|
| **5** | Feature Importance | 5 |
| **6** | Principal Component Analysis | 1 |
| **7** | Manual | 2 |
| **8** | Anova | 1 |
| **9** | Weight Score | 2 |
| **Total** | | **26** |

### *3.4.6. Identified datasets with their summary*

There are 6 open datasets out of 19 extracted from the literature. The table shows major characteristics of the datasets like number of instances, publicly availability, number of times used out of 44 and if data set is open then it's URL has been placed in URL column so that readers of this article may do further research on these datasets.

*Table 11: List of datasets and characteristics*

| SR. | Data Set | Instances | Public | Usage in SLR | URL |
|---|---|---|---|---|---|
| 1 | ILPD | 583 | Yes | 22 | https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) |
| 2 | HCVD-1 | 73 | No | 1 | |
| 3 | HCVD-2 | 1385 | Yes | 1 | https://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+%28HCV%29+for+Egyptian+patients |
| 4 | HCVD-3 | 615 | Yes | 3 | https://archive.ics.uci.edu/ml/datasets/HCV+data?fbclid=IwAR3ap0YM2IfvSeBJGe7LRj |
| 5 | HCVD-4 | 100 | No | 1 | |
| 6 | FLD-1 | 517 | No | 1 | |
| 7 | FLD-2 | 3419 | No | 1 | |
| 8 | FLD-3 | 45525 | No | 1 | |
| 9 | HCC-1 | 4423 | No | 1 | |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 10 | HCC-2 | 192 | No | 1 |  |
| 11 | HCC-3 | 139 | No | 1 |  |
| 12 | NFLD-1 | 10508 | No | 1 |  |
| 13 | NHANES-III | 12719 | Yes | 1 | https://wwwn.cdc.gov/nchs/data/nhanes3/34a/HGUHS.htm |
| 14 | LOOCV | 4032 | No | 1 |  |
| 15 | LDD-1 | 345 | Yes | 3 | https://archive.ics.uci.edu/ml/datasets/liver+disorders |
| 16 | cACLD | 497 | No | 1 |  |
| 17 | ELP-Cancer | 1541 | Yes | 1 | https://archive.ics.uci.edu/ml/datasets/Hepatitis+C+Virus+%28HCV%29+for+Egyptian+patients |
| 18 | ELP Fibrosis | 39567 | No | 1 |  |
| 19 | CPD | 186 | No | 1 |  |
| **Total** | **19 data sets** |  | **6 yes** | **44** |  |

# SLR APPENDIX A. STUDY QUALITY ANALYSIS.

*Table 12: Study Quality Score Table*

| Study ID | Reference | SQ1 Score | SQ2 Score | SQ3 Score | SQ4 Score | SQ5 Score | Total Score | Percentage N = 5 |
|---|---|---|---|---|---|---|---|---|
| 1 | Auxilia [19] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 2 | Idris [35] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 3 | Sontakke [36] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 4 | Singh [20] | 1 | 1 | 1 | 1 | 0.5 | 4.5 | 90 |
| 5 | Babu [37] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 6 | Azam [38] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 7 | Pasha [39] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 8 | Bihter [40] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 9 | Choudhary[21] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 10 | Gupta[41] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 11 | Geetha [42] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12 | Kuzhippallil[43] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 13 | Ambesange[44] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 14 | Adil[45] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 15 | Sokoliuk[46] | 1 | 1 | 1 | 1 | 0.5 | 4.5 | 90 |
| 16 | Singh [47] | 1 | 1 | 1 | 1 | 0.5 | 4.5 | 90 |
| 17 | Hartatik[22] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 18 | Shobana[23] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 19 | Saba [48] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 20 | Ambesange[49] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 21 | Syafa'ah [24] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 22 | Ahammed [50] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 23 | Mostafa [25] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 24 | Gupta [26] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 25 | Chicco [27] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 26 | Heba [28] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 27 | WU C [51] | 1 | 1 | 1 | 1 | 0.5 | 4.5 | 90 |
| 28 | Pei [52] | 1 | 1 | 1 | 1 | 0.5 | 4.5 | 90 |
| 29 | Chen[53] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 30 | Hashem[29] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 31 | Wibowo[54] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 32 | Zhaoyang [55] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 33 | Ma [30] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 34 | Deo[31] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 35 | Che [56] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 36 | Naseem [57] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 37 | Islam [58] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 38 | Kumar[59] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 39 | Vyshali [60] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |
| 40 | Haque[61] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 41 | Agarwal [32] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 42 | Esra [33] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 43 | Hashem [34] | 1 | 1 | 1 | 1 | 1 | 5 | 100 |
| 44 | Zhenbing [62] | 1 | 1 | 1 | 0 | 1 | 4 | 80 |

# SLR APPENDIX B: RESULTS OF SLR SUMMARY

*Table 13: Summary of SLR Results*

| **Author** | **Year** | **Data set** | **Data Imbalance** | **Feature selection** | **Features** | **Algorithm** | **Validation** | **Accuracy** |
|---|---|---|---|---|---|---|---|---|
| **Indian Liver Patients** | | | | | | | | |
| Auxilia [19] | 2018 | ILPD | - | Manual | 8 | Decision Tree | - | 81% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Idris [35] | 2019 | ILPD | - | No | All | Adaboost of logistic regression | - | 74.36% |
| Sontakke [36] | 2017 | ILPD | - | No | All | Artificial Neural Network | - | 73.2% |
| Singh [20] | 2020 | ILPD | - | Correlation Based | 5 | Logistic Regression | 10 fold cross validation | 74.36% |
| Babu [37] | 2017 | ILPD | - | No | All | Supervised K-means Clustering | - | 64.28% |
| Azam [38] | 2020 | ILPD | - | Correlation Based | Not Mentioned | K nearest neighbors | - | 74% |
| Pasha [39] | 2017 | ILPD | - | No | All | Grading Meta Learning | 10 fold cross validation | 71.35% |
| Bihter [40] | 2020 | ILPD | - | HP (High Performance) variable selection | Not Mentioned | Artificial Neural Network | Validation done but technique Not Mentioned | 74.14% |
| Choudhary [21] | 2021 | ILPD | - | Correlation based | All | Logistic Regression | 10 fold cross validation | 70.54% |
| Gupta[41] | 2020 | ILPD | Means Of data duplication (increasd records) | No | All | Combined Algorithm KNN, DT & ANN | - | 94.13% |
| Geetha [42] | 2021 | ILPD | - | No | All | SVM | - | 75.04% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Kuzhippallil[43] | 2020 | ILPD | SMOTE | Combination of GA & XGBoost classifier results | Not Mentioned | Stacking Estimator | - | 85% |
| Ambesange[44] | 2020 | ILPD | - | Correlation Matrix | Not Mentioned | KNN Model | - | 91% |
| Adil[45] | 2018 | ILPD | - | No | All | Logistic Regression | 5 fold cross validation | 74% |
| Sokoliuk[46] | 2020 | ILPD | - | No | All | KNN | 5 fold cross validation | 74% |
| Singh [47] | 2018 | ILPD | - | - | - | KNN | 10 fold cross validation | 73.7% |
| Hartatik[22] | 2020 | ILPD | - | Feature Correlation | 6 | Naïve Bayes | - | 72.50% |
| Shobana[23] | 2021 | ILPD | - | Feature Correlation | 5 | Gradient Boosting Algorithm | - | 94% |
| Saba [48] | 2016 | ILPD | Use of multiple layers | F-score feature selection | All initially | KNN enabled HM-BAGMOOV | 10 fold cross validation | 72.7% |
| Ambesange[49] | 2020 | ILPD | SMOTE | Feature Importance | Not Mentioned | Random Forest | - | 74% |
| **HCV Dataset (HCVD)** | | | | | | | | |
| Syafa’ah [24] | 2021 | HCVD-1 | - | No | 10 | Random Forest | - | 90% |
| Ahammed [50] | 2020 | HCVD-2 | Synthetic minority oversampling | Importance | - | K-NN | 10 fold cross validation | 94% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | technique | | | | | |
| Mostafa [25] | 2021 | HCVD-3 | Stratified Sampling (3:1) | Principal Component Analysis | 4 | SVM | 10 fold cross validation | 98.23 % |
| Gupta [26] | 2021 | HCVD-3 | Random sampling (80:20) | Manual | 4 | Random Forest | - | 94.33 % |
| Chicco [27] | 2021 | HCVD-3 | - | Mann-Whitney U & Chi square based correlation | 9 for Mann-Whitney<br>3 for chi square | Random Forest | - | 97.1 % |
| Heba [28] | 2015 | HCVD-4 | - | Significance based on ANOVA | 9 | Decision Tree | - | 93.70 % |
| **Fatty Liver Disease (FLD)** | | | | | | | | |
| WU C [51] | 2018 | FLD-1 | - | Weight score based | 10 | Random Forest | 10 fold cross validation | 86.48 % |
| Pei [52] | 2020 | FLD-2 | Over Sampling | Feature importance | 12 | XG-Boost classification | 10 fold cross validation | 94.15 % |
| Chen[53] | 2018 | FLD-3 | - | No | ALL | Multi layer random forest | - | 98.63 % |
| **Hepatocellular (HCC)** | | | | | | | | |
| Hashem[29] | 2020 | HCC-1 | - | Corelation based | 5 | ADtree | 10 fold cross validation | 95.6 % |
| Wibowo[54] | 2020 | HCC-2 | - | - | All | Random Forest | - | 100 % |
| Zhaoyang [55] | 2020 | HCC-3 | | Feature Importance | All | Random Forest | 10 fold cross validation | 90% AUC based |
| **Non Fatty Liver Disease (NAFLD)** | | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ma [30] | 2018 | NAFLD-1 | - | Weight score | 5 | Bayesian Network | 10 fold cross validation | 83% |
| Deo[31] | 2019 | NHANES-III | SMOTE | - | 11 | Boosted Tree model | 10 fold cross validation | 79 % |
| **LOOCV** | | | | | | | | |
| Che [56] | 2021 | LOOCV | - | - | All | Convolutional neural network | 10 fold cross validation | 90% |
| **Liver Disorder Datset (LDD)** | | | | | | | | |
| Naseem [57] | 2020 | LDD-1 | Bootstrap Sampling | - | 7 | Random Forest | 10 fold cross validation | 72.17 % |
| Islam [58] | 2019 | LDD-1 | - | No | All | Naïve Bayes | - | 96.52 % |
| Kumar[59] | 2018 | LDD-2 | - | No | All | C5.0 with Adaptive Boosting | 10 fold cross validation | 75.19 % |
| Vyshali [60] | 2018 | LDD-3 | - | No | All | Linear Discrimenent | - | 95.8 % |
| Saba[48] | 2016 | LDD-1 | Use of multiple layers | F-score feature selection | All initially | KNN enabled HM-BAGMOOV | 10 fold cross validation | 70.16 % |
| Haque[61] | 2018 | LDD-1 | - | Correlation based | 6 | Artificial Neural Network | 10 fold cross validation | 85.29 % |
| **Compensated Advanced Chronic Liver Disease (cACLD)** | | | | | | | | |
| Agarwal [32] | 2021 | cACLD | - | Manual based on domain knowledge | 12 | XGboost | 10 fold cross validation | 98.7 % |
| **Egyptian liver patients (ELP)** | | | | | | | | |
| Esra [33] | 2019 | ELP-cancer | - | Correlational | 17 | Random forest | 5 fold cross | 86.96 % |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | Attribute Evaluation | | | validation | |
| Hashem [34] | 2018 | ELP-Fibrosis | - | Correlation Method | 4 | Alternating Decision Tree | 10 fold cross validation | 84.40 % |
| **Cirrhosis Patient Data (CPD)** | | | | | | | | |
| Zhenbing [62] | 2020 | CPD | - | Feature Importance | 4 | BP Neural Network | - | 80.4 % |

### Chapter Summary

In this chapter literature related to domain has been analyzed in detail. A systematic and protocol literature review has been done. Answers to the research question has been given with statistical analysis. The results of this chapter are assisting in preparing the framework for identified data set. The framework needs to be effective, accurate and unified that can handle multiple data sets one by one for results. 44 articles have been extracted and reviewed from 5 different electronic repositories published from January 2015 to November 2021. After a systematic and protocol-based review we answered 6 research questions about machine learning algorithms. The identification of effective feature selection technique, data imbalance management technique, accurate machine learning algorithms, list of available data sets with their URL's and characteristics, and feature importance based on usage has been identified for diagnosing the liver disease. The reason to selecting this research question is, in any machine learning framework the role of dimensionality reduction, data imbalance management, Machine learning Algorithm with its accuracy and data itself is very significant.

CHAPTER 4

# METHODOLOGY

## 4.1. MACHINE LEARNING

A branch of artificial intelligence, that is used to analyze data and learn from it. Based on that learning pattern may identified and relevant decision can be taken with minimal human intervention. Machine learning is although field of computer science, but it is totally different from tradition computing technology. The aim is to make machine intelligent enough that it can learn from environment and remain adaptive. Traditional algorithms are programmed with static behavior in any circumstances, but machine learning algorithms takes dataset as input based on which it gets trained. The training helps in generating predicted outcomes for any new data instances. There are many areas in computing that are being benefitted by machine learning for example facial recognition by traditional social media network, whenever somewhere in world someone upload a picture that may contain you in it the social media immediately informs you whether it is you. Similarly, pattern recognition, signal processing, ecommerce recommendation systems and many more as shown in Figure 15: Application of Machine Learning, all using machine learning to make adaptive and accurate decisions. There are three major types of learning in Machine Learning 1) Supervised Learning 2) Unsupervised Learning 3) Semi-Supervised Learning. All three types have been discussed further in following section.

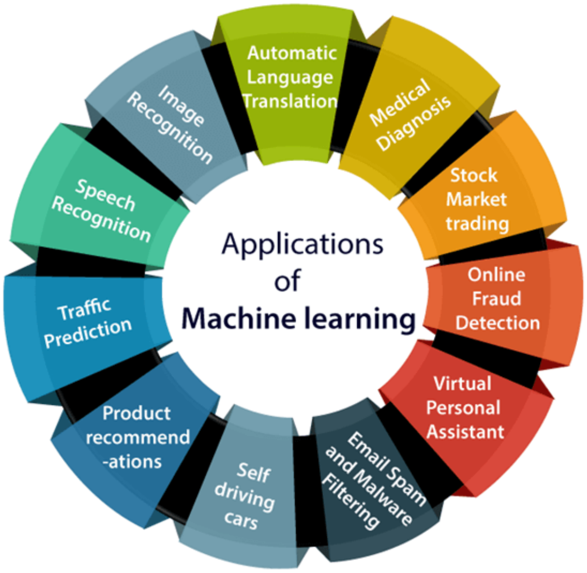


*Figure 15: Application of Machine Learning*

### *4.1.1. Supervised Learning*

In supervised learning, the algorithms required an input dataset that is labelled based on which it trains the computer. After completion of training, for testing purpose some testing data given to computer by removing its labels and checked whether it gives accurate results or not. Based on testing results the accuracy of an algorithm is extracted. For example, in supervised learning a dataset given to computer of pencils and erasers based on which computer starts training. Computer after finishing the learning of patterns and shapes of all given samples starts predicting new unlabeled samples. The most common use of supervised learning is to analyze the historical patterns in data and make effective decisions, currently used by most of the organizations. A brief working of supervised learning algorithms is shown in Figure 16: Overview of Supervised Learning.

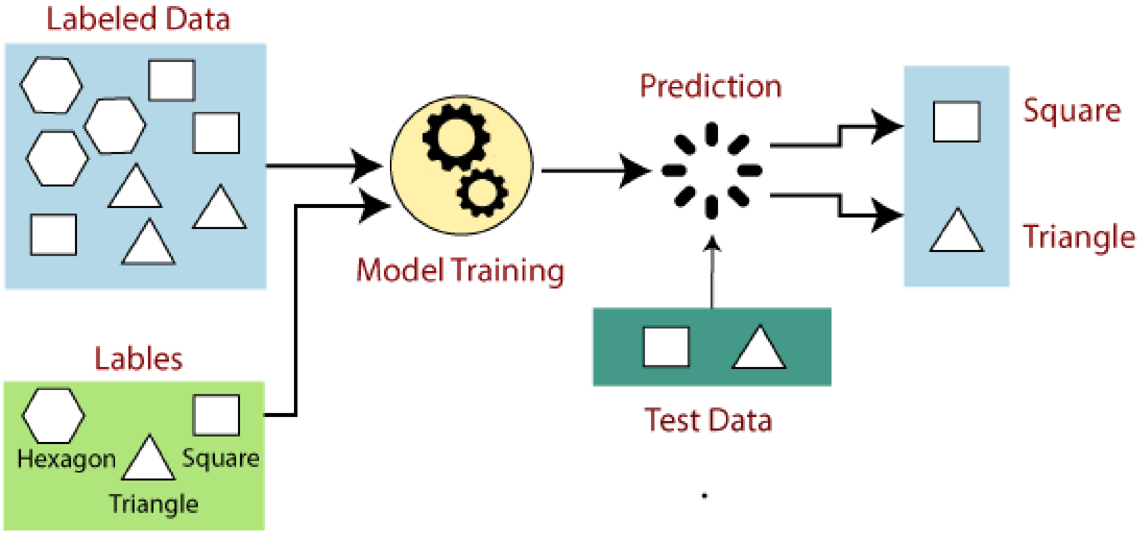


*Figure 16: Overview of Supervised Learning*

### *4.1.2. Unsupervised Learning*

The unsupervised learning uses Artificial Intelligence algorithms to recognize or identify patterns from data that is neither labeled nor classified. Target is to identify similar type of data instances and grouping it. Now it depends on user that how many groups are required. The specification of number of groups is another research gap. The accuracy of unsupervised learning algorithm is also dependent on number of groups, although there are other many factors like algorithm that is being used. Characteristics of data that is being used and many more. Commonly used algorithms for unsupervised learning are categorized under clustering, Association. In Figure 17: Unsupervised Learning Illustration shown to validate above description. The data is unlabeled number of clusters or groups should be 2 to achieve the required result. But if we change number of clusters other than 2 it cannot produce accurate results. As there are two types of items as input data. What if number of clusters considered as 3 or number of clusters given 4. So the case study validating the importance of setting value of K in K number of clusters.

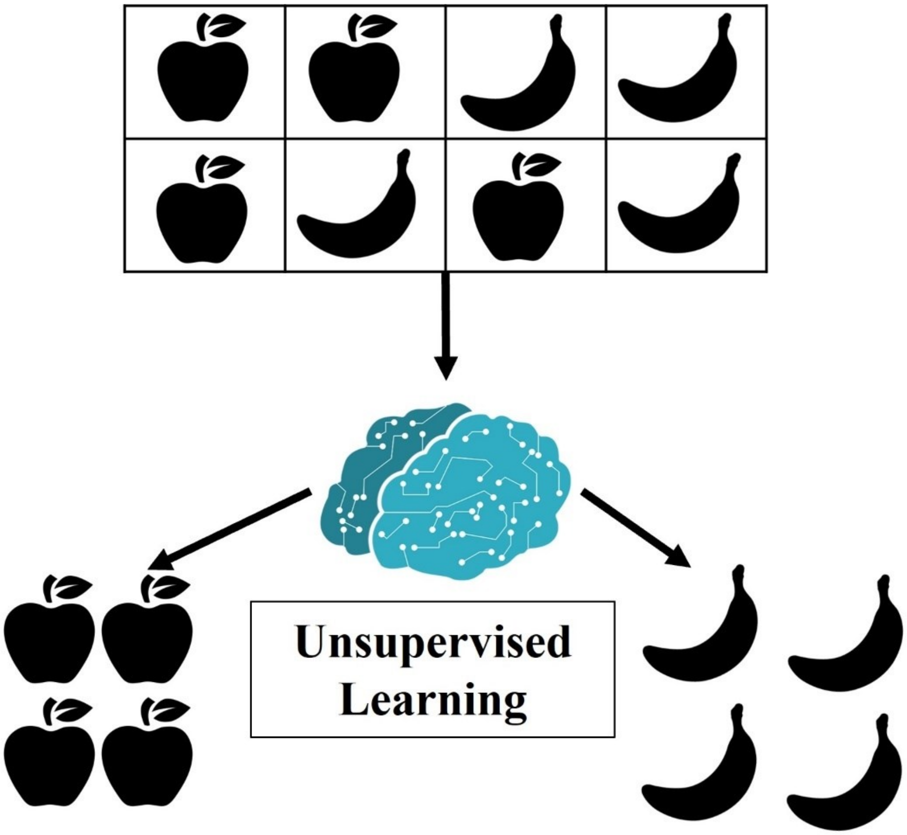


*Figure 17: Unsupervised Learning Illustration*

### *4.1.3. Semi-Supervised Learning*

In semi-supervised learning algorithms, there are some instances in input data that are labelled, based on which unlabeled data may obtain labels. There are three major assumptions considered in semi-supervised algorithms. Continuity Assumption, Manifold Assumption, Cluster Assumption. In continuity assumptions the machine learning algorithm assumes that the point which closer to other are likely to have same output. For example, a data instance is closest to be an apple the output will be apple. In manifold assumption the distances and densities features are used called manifold and in cluster assumption the whole data instances divided in given discrete number of clusters that are more likely to share same output labels. Application of semi-supervised learning includes speech analysis to label audio files, protein sequence classification to label DNA strands and Content on internet classification for labeling of webpages.

## 4.2. APPLIED ALGORITHMS

In this research activity based on SLR results 6 most effective algorithm has been identified in domain of prediction of liver disease. Algorithms are Decision Tree, Logistic Regression, Deep Learning, K-Nearest Neighbors, Naïve Bayes, and Random Forest. All mentioned algorithms are targeting the classification of data instances. Supervised learning has been used for classification of data instances as all datasets are labelled.

### *4.2.1. Decision Tree*

Decision tree is one of the supervised learning algorithms that is being used for classification of data instances. Like other algorithms it can also be used for regression problems. It aims to classify data input instances in relent class based on fixed decision rules, the rules made based on training from the training data instances. Once a tree based on rules has been made, every instance starts getting tested from the root node. Based on the value the branches of tree to be followed one by one to reach the matching label. There are two major continuous variable decision trees with continuous target variables and categorical variable decision tree based on categorical target variables.

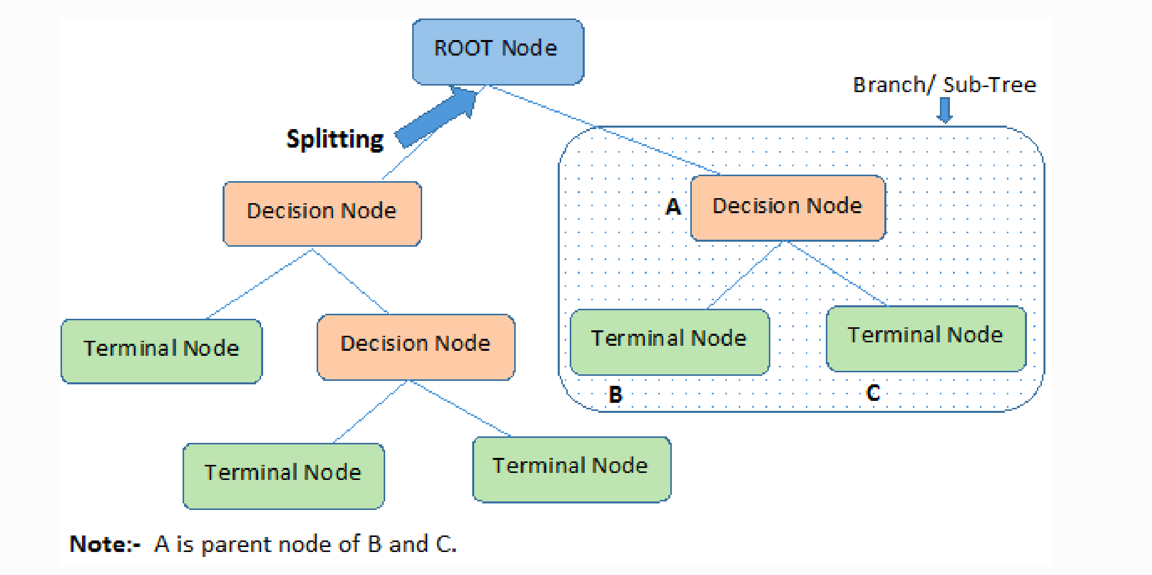


*Figure 18: Description of ingredients in Decision Tree by Kdnuggets*

In Figure 18: Description of ingredients in Decision Tree by Kdnuggets showing the basic terminologies of decision tree. The root node to represent the entire population that may further divided in subsets, the splitting node to divide a particular node into two or more sub nodes, the decision node to show further splitting of a certain node into sub nodes, the leaf or terminal node that do not split further, the pruning refers to removal of sub-node of a decision node, the branch also called sub tree to represent a subsection of an entire tree, parent and child node in which the node that is being splitted is called parent and the resultant nodes are called child nodes.

Steps taken to create a decision tree using entropy.

- Step 1 includes calculation of entropy of the target.

$$Entropy\ (Target) = Entropy(P, N)$$
$$using\ Formula$$

$\mathrm{E}E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$

- Step 2 includes splitting of data set in different attributes. For each branch entropy will be calculated and tree further divided. Information gain calculated using the formula

$$Gain(P, N) = Entropy(P) - Entropy(P, N)$$

- Step3 includes choosing attribute with largest information gain as a decision node.
- Step 4 the nodes with entropy = 0 is considered as leaf node and it won't split further but attribute with entropy other than 0 will be splitted further.
- Step 5 the above-mentioned steps will be called recursively until all leaf nodes may identified with labels.

Following parameters has been used while using decision tree in this research as shown in Table 14: Decision Tree Parameters.

*Table 14: Decision Tree Parameters*

| Sr. | Parameter | Value |
|---|---|---|
| 1 | Criterion | Gain ratio |
| 2 | Maximal Depth | 10 |
| 3 | Pruning | True |
| 4 | Pre-Pruning | True |
| 5 | Confidence | 0.1 |
| 6 | Minimal gain | 0.01 |
| 7 | Minimal Leaf Size | 2 |
| 8 | Minimal size for split | 4 |
| 9 | Pre-pruning alternatives | 4 |

### *4.2.2. Logistic Regression*

Logistic regression belongs to family of supervised learning algorithms or classification algorithms. It is used to assign observation to a discrete set of classes. For example, a person with Covid positive or not, an email is spam or not, person is male or female, a transaction is fraud activity or not and many more. Two main types include binary logistic regression and multi-linear functions fails class. In this research binary logistic regression has been used to classify liver datasets. The logistic regression can be referred as linear regression but the function that is being used by logistic regression is complex and named as 'Sigmoid Function'. The sigmoid function is used to map the real value into any other value ranging 0 to 1.

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

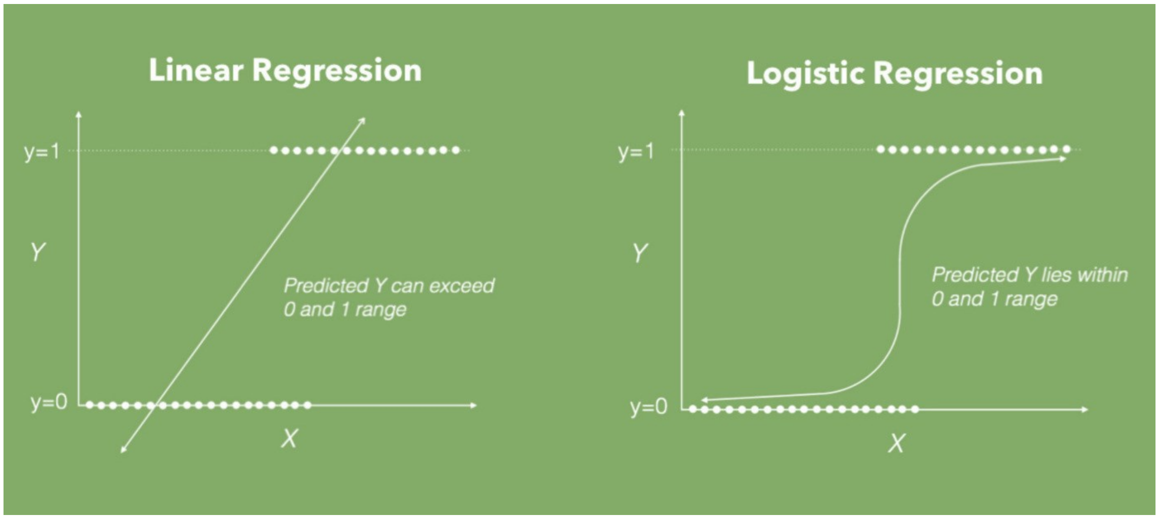


*Figure 19: Linear Regression VS Logistic Regression Graph| Image: Data Camp*

The Figure 19: Linear Regression VS Logistic Regression Graph| Image: Data Camp shows a brief difference between linear regression and logistic regression. Parameters used for logistic regression has been given in Table 15: Parameters used for logistic regression

*Table 15: Parameters used for logistic regression*

| Sr. | Parameter | Value |
|---|---|---|
| 1 | Solver | AUTO |
| 2 | Standardize numeric column | True |
| 3 | Add intercept | True |
| 4 | Computer P Values | True |
| 5 | Remove Collinear columns | True |
| 6 | Missing value handling | Mean Imputation |
| 7 | Max iterations | Default |
| 8 | Max runtime | Default |

### *4.2.3. Deep learning*

The deep learning is a branch of machine learning as shown in Figure 20: Deep Learning, a neural network with more than or equal to 3 layers. The neural network tried to simulate the behavior same as human brain, that allows to learn the computer from large amount of data set. Eventually the learning pattern may help in recognizing or classifying the new instances. Some time the predicted outputs contradict the expectation, in this scenario some layers can be added to enhance the output. This is not a rule of thumb that you must

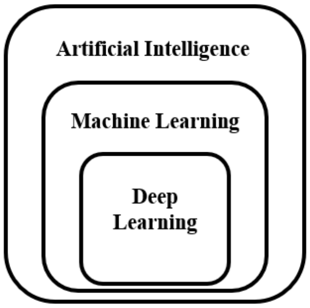


*Figure 20: Deep Learning*

add always, may be sometimes you have to reduce number of layers from the neural network. The heart of deep learning is neural network. It was designed aiming to mimic the working of human brain. The components of Neuron are shown in Figure 21:

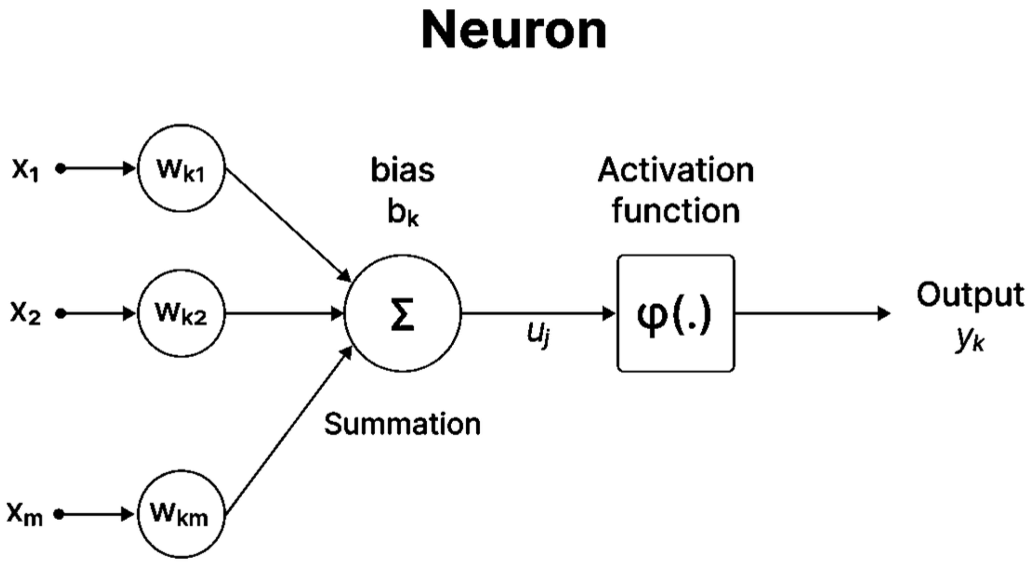


*Figure 21: Components of Neuron by v7labs*

Components of Neuron by v7labs. There are two motivtions for neuronal perception of deep learning. Number one there is an assumption that intelligent behaviour is possible as proved by human brain. Based on reverse engineering there is a possibility of building an intelligent system. Secondly to match the human brain intelligence there is a need of making a mathematical model that could answer the fundamental scientific questions. In this research the parameters used are shown in Table 16: Deep Learning used Parameters.

*Table 16: Deep Learning used Parameters*

| Sr. | Parameter | Value |
|---|---|---|
| 1 | Activation | Rectifier |
| 2 | Hidden layer sizes | 50,50 |
| 3 | Train samples per iteration | -2 |
| 4 | Adaptive rate | True |
| 5 | epsilon | 1.0E-8 |
| 6 | rho | 0.99 |
| 7 | Standardize | True |
| 8 | L1 | 1.0E-5 |
| 9 | L2 | 0.0 |
| 10 | Max w2 | 10.0 |
| 11 | Loss function | Auto |
| 12 | Distribution function | Auto |

An illustration of a neural network with one hidden layer has been shown in Figure 22: Neural Network of 1 hidden layer.

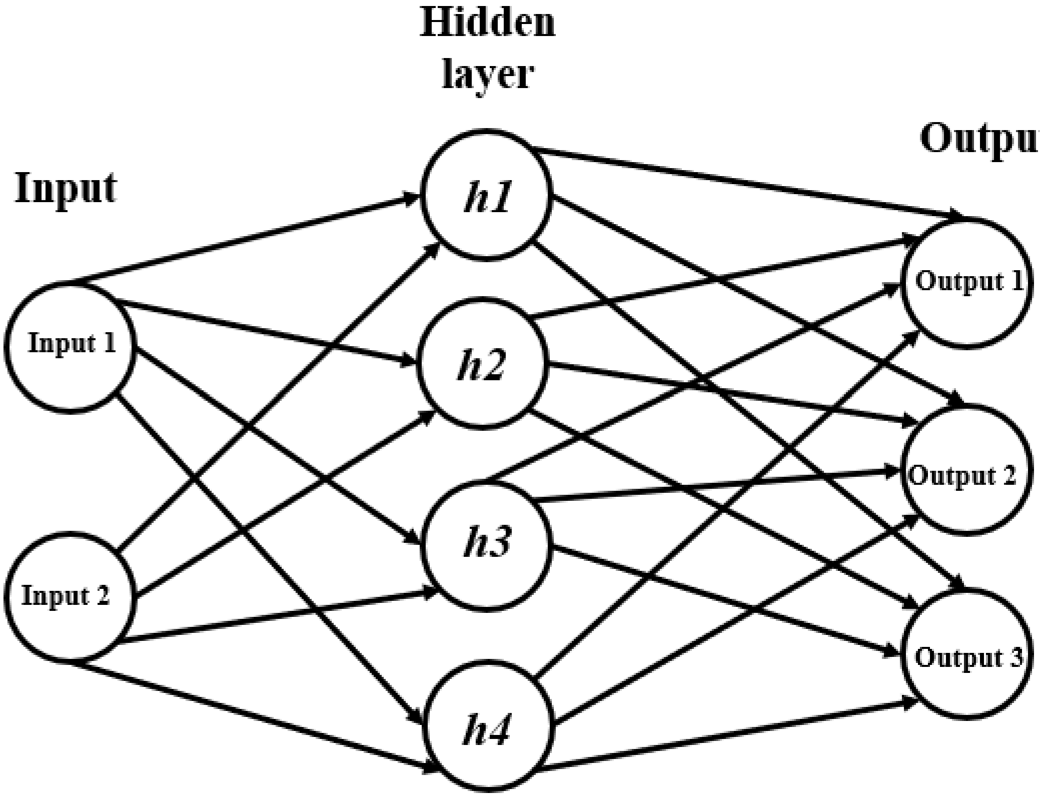


*Figure 22: Neural Network of 1 hidden layer*

### *4.2.4. K nearest neighbours*

A type of supervised learning in which a data instance classified based on its neighbours. For example there is a new data instance, algorithm will cross match it with k nearest neighbours. There is a need of specifying the value of k in order to match with k number of neighbours. Neighbours will help in classifying the new data instance. If there are more neighbours of class A as compared to class B then the new instance will be predicted as of class A. there is critical thing in K nearest neighbour algorithm. The value of K, it is the most difficult to assign the value of K, one of the key factor that has a strong influence on the algorithm. In Figure 23: K nearest neighbor the problem of k nearest neighbour has been shown consider the star shaped data instance as a new one, if we assign k = 3 the new instance will be labelled as a square but if we assign 5 as value of k then new instance will be classified as a circle as 3 neighbors are circle and 2 are square whereas in case of k = 3 there are two squares and one circle. This problem is itself a research topic, and many researchers already working on it.

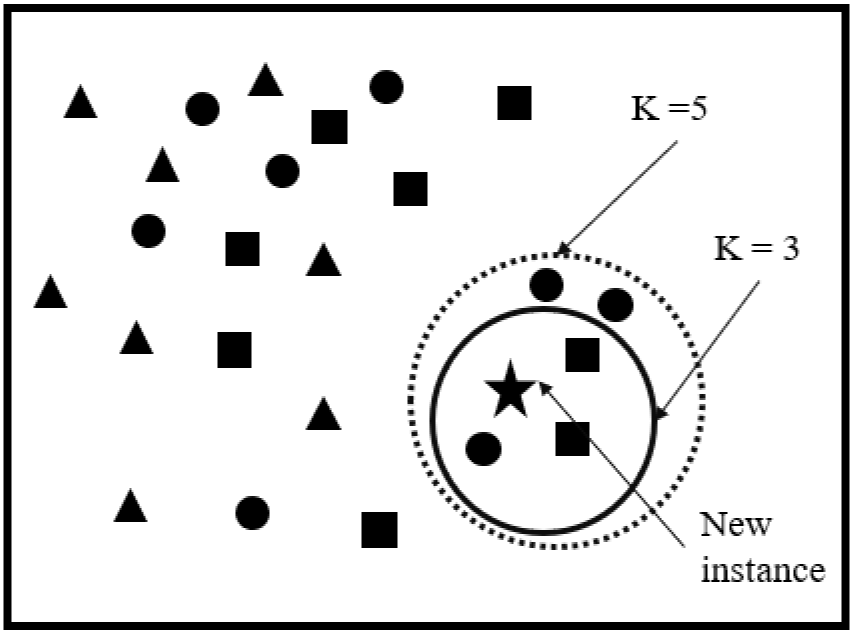


*Figure 23: K nearest neighbor*

In this research activity the parameters that has been used are shown in Table 17: Parameters used for K Nearest Neighbors. Reason to set value of k = 9 is accuracy. As on k = 9 model was outperforming as compared to other values.

*Table 17: Parameters used for K Nearest Neighbors*

| Sr. | Parameter | Value |
|---|---|---|
| 1 | Value of K | 9 |
| 2 | Weighted Vote | True |
| 3 | Measure Types | Mixed Measures |
| 4 | Mixed Measure | Mixed Euclidean Distance |

### 4.2.5. Naïve bayes classifier

The foundation of Naïve Bayes algorithm is based on probabilistic model for classification. There is pivotal role of Bayes theorem as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With the help of Bayesian theorem, probability of happening of A, given that B has happened can be calculated. In this case A is the hypothesis and B is the evidence. It is assumed that the predictors are not dependent. It means that the presence of a feature doesn't affect the other feature. Also refers as naïve.

$$\text{For} \qquad B = (b_1, b_2, b_3, \dots \dots, b_n)$$

Where B is data set and the $b_1, b_2, b_3, \dots \dots, b_n$ are instances of data sets.

The Bayesian theorem for B dataset will be

$$P(A|b_1, b_2, b_3, \dots \dots, b_n) = \frac{P(b_1|A)P(b_2|A) \dots P(b_n|A)P(A)}{P(b_1)P(b_2) \dots . P(b_n)}$$

There are three main types of naïve bayes classifier

1) Multinomial Naïve Bayes
   mostly used for document classification.
2) Bernoulli Naïve Bayes
   same as multinomial classifier but result is in binary. For example, a document belongs to news category or not.
3) Gaussian Naïve Bayes
   Use gaussian distribution, for the continuous values.

In this research naïve bayes with Laplace correlation has been used.

### *4.2.6. Random Forest*

When large number of decision trees get together as an ensemble of classifier the random forest formed. Each decision tree predict output. Based on algorithm of ensemble being

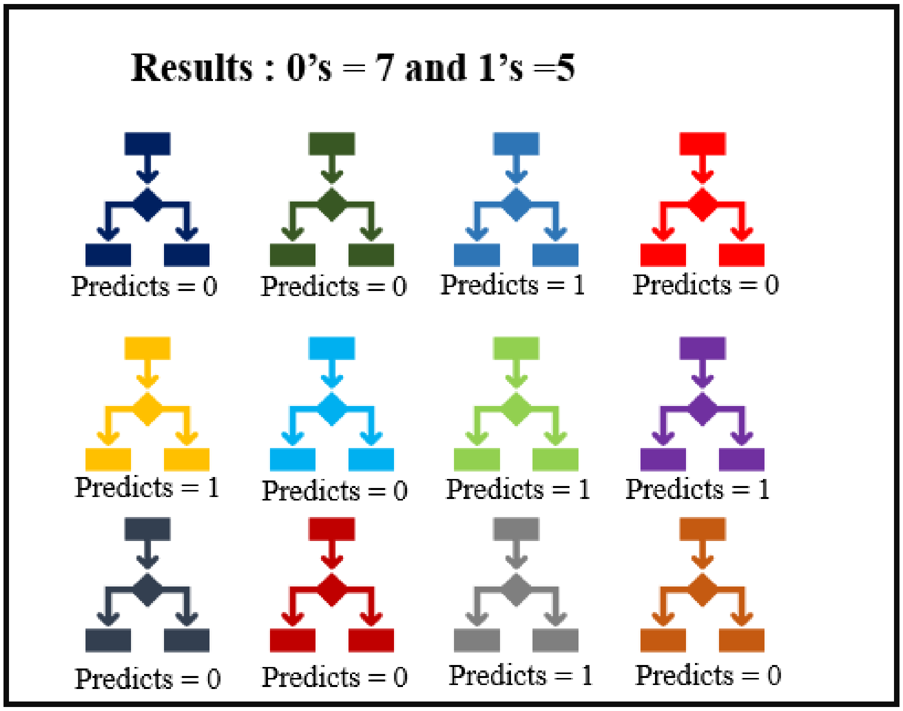


*Figure 24: Random Forest Algorithm*

used in decision tree the final output formed. For example, there are 12 decision trees ensembled together as shown in Figure 24: Random Forest Algorithm there are 7 decision trees generating output = 0 and 5 that generating output = 1 the maximum numbers saying it's 1 so that the final output will be 1.

The parameters used in this research has been listed as Table 18: Parameters used for Random Forest.

*Table 18: Parameters used for Random Forest*

| Sr. | Parameter | Value |
|---|---|---|
| 1 | Number of Trees | 100 |
| 2 | Criterion | Gain Ratio |
| 3 | Maximal Depth | 10 |
| 4 | Guess subset ratio | True |
| 5 | Voting Strategy | Confidence Vote |
| 8 | Parallel Execution | True |

## 4.3. PROPOSED FRAMEWORK (MaLLiDD)

The proposed framework **Ma**chine **L**earning Based **Li**ver **D**isease **D**iagnose (MaLLiDD) has been shown as Figure 25: MaLLiDD the proposed framework.

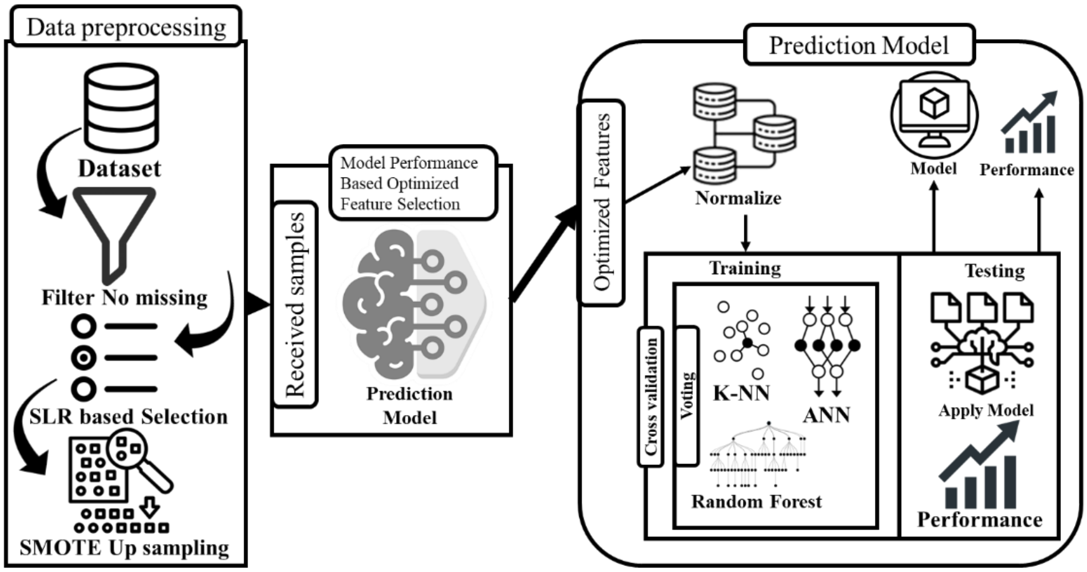


*Figure 25: MaLLiDD the proposed framework*

The brief explanation of ingredients in MaLLiDD as follow.

### *4.3.1. Filter Examples*

An operator used to filter out the attributes with missing values. The goal is to only use data that doesn't have any missing value. There are other ways to handle missing values like replacing missing value with average or imputation. But due to sensitivity of application of framework only data that is labelled has been used.

### *4.3.2. SLR based attribute selection*

Feature importance has been shown in Table 8: Table of Identified important features and Figure 12:Feature Usage Graph F1 to F22. The listed features were used as an experiment, based on performance the manual selection may change one by one. For example, there are 22 important identified features from all 3 datasets. So before applying the MaLLiDD first look for important features in input dataset. Select the most important one based on list of Table 8 so that run time of optimized attribute selection may get reduced.

### *4.3.3. SMOTE Upsampling*

The SMOTE stands for Synthetic Minority Oversampling Technique used to tackle class imbalance problem. In this approach the samples of minority class get duplicated to make the classification a balanced one. This is also referred as type of data augmentation for the minority class. In SMOTE the selection of examples has done that are close in feature space, A line drawn between examples and new samples drawn along that line. "SMOTE first selects a minority class instance a at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b" [63].

The parameters for SMOTE Upsampling that has been used in this research are listed as Table 19: SMOTE Upsampling parameters

*Table 19: SMOTE Upsampling parameters*

| Sr. | Parameter | Value |
|---|---|---|
| 1 | Number of neighbors | 5 |
| 2 | Equalize classes | True |
| 3 | Auto detect minority class | True |
| 4 | Nominal change rate | 0.5 |
| 5 | Use local random seed | False |

### *4.3.4. Brute Force optimized feature selection*

The operator Optimized Selection (brute force) is a nested kind of operator. Inside the operator you must specify your model for which you need high performing features. In our case the whole prediction model that has been shown in Figure 25: MaLLiDD the proposed framework. The third part of framework is prediction model and used in brute force feature selection. The process brute force selection starts iterating unless the most relevant feature found. The selected features then assigned by weights = 1 and delivered as output of the optimized features brute force. The minimum number of attributes were set to 1 and weight normalization were checked true while using it in current research.

### *4.3.5. Normalize*

After brute force feature selection, the outperforming features given to normalize as input. To normalize the feature "proportion transformation method" has been used.

### *4.3.6. Cross Validation*

The cross validation applied to check the statistical performance of the learning model. The learning model is a subprocess of the operator cross validation. The two subprocesses in cross validation are training and testing process. While working, cross validation operator split the example set or dataset in k subsets of equal size. Of all the k subsets there is a one subset that retained as a test dataset and the rest all dataset subsets used as training dataset. The cross-validation process then repeated k times for each of the k subsets. Then the results of k iteration got averaged and results given as an output. With the help of cross validation, it is easy to identify under or over fitting of model. There are 20 numbers of folds used for cross validation, the sampling type is "Stratified", and parallel execution has been enabled.

### *4.3.7. Voting the ensemble*

The voting ensemble is a nested process in which at least two learning algorithms applied for classification. It uses majority vote for classification or the average for regression case. For example, if 2 leaning model predicts output for a data instance as 1 and 1 algorithm predicts it as 0 then the voting ensemble will declare as 1. As there are two votes for output = 1 which is more than 1. A short example already shown in Figure 24: Random Forest Algorithm. A theoretical

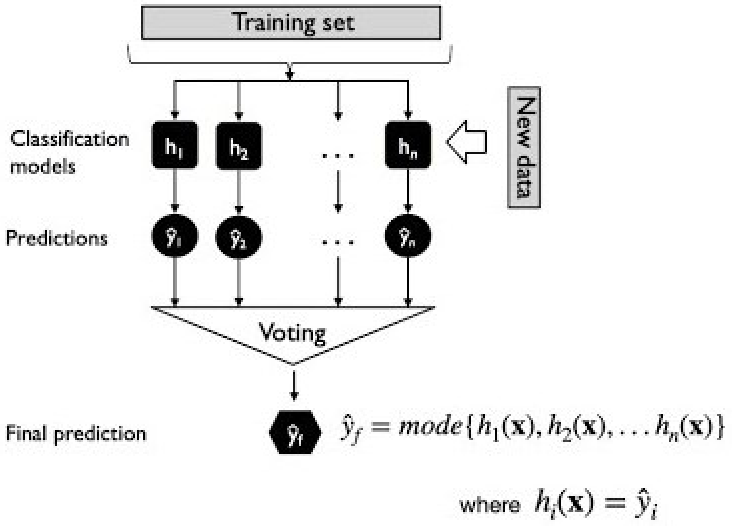


*Figure 26: Voting the classifier*

explanation also given in Figure 26: Voting the classifier. For voting ensemble in this research activity random forest, Artificial neural network and K-nearest neighbors has been used. The detailed explanation of the random forest has already been given in current chapter under 4.2.6. Random Forest, ANN under 4.2.3. Deep learning and KNN under 4.2.4. K nearest neighbours .

### *4.3.8. Apply Model*

The operator used to apply a model on dataset. The model in first phase gets trained with the help of an input dataset that can be further used for unlabeled or testing data set.

### *4.3.9. Performance Binomial*

The operator performance binomial is used to evaluate a learning model statistically. With the help of this operator following values has been obtained.

- ✓ Accuracy
- ✓ Classification Error
- ✓ Precision
- ✓ Recall
- ✓ F measure
- ✓ Sensitivity
- ✓ Specificity

These all parameters helped us in evaluating the results.

## 4.4. Overview of Application of MaLLiDD

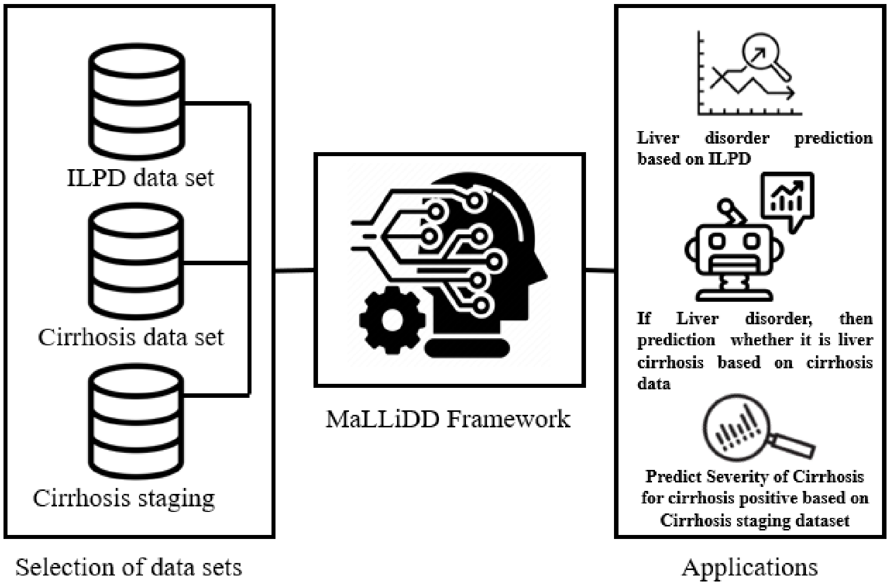

ILPD data set
Cirrhosis data set
Cirrhosis staging
Selection of data sets
MaLLiDD Framework
Liver disorder prediction based on ILPD
If Liver disorder, then prediction whether it is liver cirrhosis based on cirrhosis data
Predict Severity of Cirrhosis for cirrhosis positive based on Cirrhosis staging dataset
Applications

*Figure 27: Application of MaLLiDD*

The application of the framework MaLLiDD has been illustrated as Figure 27: Application of MaLLiDD. First of all one has to select a dataset based on required results for example if someone want to diagnose either he / she has a liver disorder he / she will select ILPD data set and test by entering the required values. If the liver disorder detected, it'll be classified as a liver cirrhosis positive or negative and if positive it'll be further classified as severe or not.

**Chapter Summary**

In this chapter details of proposed framework have been discussed with its reason to use. All the preliminaries have been discussed so that thesis may be more expressible for the readers. The further research gaps have been also identified in parallel areas of improvement also discussed, so if one can change approach for better results. The algorithms theory with it's working has been discussed. All operator with their working discussed with usage details including parameters values.

CHAPTER 5

# EXPERIMENTAL EVALUATION

The section consists of results and evaluation of methodology.

## 5.1. USED DATASETS

There are 3 data sets that has been used for experimental evaluation. The proposed methodology has been tested on three data sets. Discussed in detail as sections given below.

### *5.1.1. Cirrhosis data set*

The statistical description of dataset is shown in Table 20: Cirrhosis Data set description

*Table 20: Cirrhosis Data set description*

| Sr. | Name | Type | Missing | Description |
|---|---|---|---|---|
| 1 | Category | Binomial | 0 | Positive = 1 (030)<br>negative = 0 (585) |
| 2 | ID | Integer | 0 | Maximum = 615 |
| 3 | Age | Binomial | 0 | Minimum = 19<br>Maximum = 77<br>Average = 47.408 |
| 4 | Sex | Binomial | 0 | Male = 377<br>Female = 238 |
| 5 | ALB | Real | 1 | Minimum = 14.900<br>Maximum = 82.200<br>Average = 41.620 |
| 6 | ALP | Real | 18 | Minimum = 11.300<br>Maximum = 416.600<br>Average = 68.284 |
| 7 | ALT | Real | 1 | Minimum = 0.900<br>Maximum = 325.300<br>Average = 28.451 |
| 8 | AST | Real | 0 | Minimum = 10.600<br>Maximum = 324<br>Average = 34.786 |
| 9 | BIL | Real | 0 | Minimum = 0.800<br>Maximum = 254<br>Average = 11.397 |
| 10 | CHE | Real | 0 | Minimum = 1.420<br>Maximum = 16.410<br>Average = 8.197 |
| 11 | Chol | Real | 10 | Minimum = 1.430 |

| | | | | |
|---|---|---|---|---|
| | | | | Maximum = 9.670<br>Average = 5.368 |
| 12 | CREA | Integer | 0 | Minimum = 8<br>Maximum = 1079<br>Average = 81.289 |
| 13 | GGT | Real | 0 | Minimum = 4.500<br>Maximum = 650.900<br>Average = 39.533 |
| 14 | PROT | Real | 1 | Minimum = 44.800<br>Maximum = 90<br>Average = 72.044 |

### *5.1.2. Indian Liver Patient Data Set (ILPD)*

The statistical description of dataset is shown in Table 21: Indian Liver Patient Data Set (ILPD)

*Table 21: Indian Liver Patient Data Set (ILPD)*

| Sr. | Name | Type | Missing | Description |
|---|---|---|---|---|
| 1 | Target | Binomial | 0 | Positive = 1 (416)<br>negative = 2 (167) |
| 2 | Age | Integer | 0 | Minimum = 4<br>Maximum = 90<br>Average = 44.746 |
| 3 | Gender | Binomial | 0 | Least = Female (142)<br>Most = Male (441) |
| 4 | Total Bilirubin | Real | 0 | Minimum = 0.400<br>Maximum = 75<br>Average = 3.299 |
| 5 | Direct Bilirubin | Real | 0 | Minimum = 0.100<br>Maximum = 19.700<br>Average = 1.486 |
| 6 | Alkaline Phosphate | Integer | 0 | Minimum = 63<br>Maximum = 2110<br>Average = 290.576 |
| 7 | Alamine Aminotransferase | Integer | 0 | Minimum = 10<br>Maximum = 2000<br>Average = 80.714 |
| 8 | Aspartate Aminotransferase | Integer | 0 | Minimum = 10<br>Maximum = 4929<br>Average = 109.911 |
| 9 | Total Proteins | Real | 0 | Minimum = 2.700<br>Maximum = 9.600<br>Average = 6.483 |
| 10 | Albumin | Real | 0 | Minimum = 0.900<br>Maximum = 5.500<br>Average = 3.142 |

| | | | | |
|---|---|---|---|---|
| 11 | A/G Ratio Albumin and Globulin Ratio | Real | 4 | Minimum = 0.300<br>Maximum = 2.800<br>Average = 0.947 |

### *5.1.3. Cirrhosis Staging Data Set*

The statistical description of Cirrhosis data set is as following. In Table 22: Cirrhosis Staging Data Set stats.

*Table 22: Cirrhosis Staging Data Set stats*

| **Sr.** | **Name** | **Type** | **Missing** | **Description** |
|---|---|---|---|---|
| 1 | Stage | Binomial | 6 | Severe =1 (299)<br>Not Severe = 0 (113) |
| 2 | ID | Integer | 0 | Unique identifier |
| 3 | N_Days | Integer | 0 | Minimum = 41<br>Maximum = 4795 |
| 4 | Status | Polynomial | 0 | CL (25)<br>C (232)<br>D (161) |
| 5 | Drug | Polynomial | 0 | D-Pencilla (158)<br>Placebo (154)<br>NA (106) |
| 6 | Age | Integer | 0 | Minimum = 9598<br>Maximum = 28650<br>Average= 18533.35<br>Deviation = 3815.84 |
| 7 | Sex | Binomial | 0 | Female (374)<br>Male (44) |
| 8 | Ascites | Binomial | 0 | Yes (24)<br>No (288)<br>NA (106) |
| 9 | Hepatomegaly | Binomial | 0 | Yes (160)<br>No (152)<br>NA (106) |
| 10 | Spiders | Binomial | 0 | Yes (90)<br>No (222)<br>NA (106) |
| 11 | Edema | Polynomial | 0 | Yes (20)<br>No (354)<br>S (44) |
| 12 | Bilirubin | Real | 0 | Min = 0.300<br>max = 28<br>Avg = 3.221<br>Deviation = 4.408 |
| 13 | Cholesterol | Integer | 0 | Min = 120 |

| | | | | |
|---|---|---|---|---|
| | | | | Max = 28<br>Average = 3.221<br>NA (134) |
| 14 | Albumin | Real | 0 | Min = 1.960<br>Max = 4.640<br>Average = 3.497 |
| 15 | Copper | Integer | 108 | Min = 4<br>Max = 588<br>Avg = 97.648 |
| 16 | Alk_Phos | Real | 106 | Min = 289<br>Max = 13862.400<br>Average = 1982.656 |
| 17 | SGOT | Real | 106 | Min = 26.350<br>Max = 457.250<br>Average = 122.556 |
| 18 | Tryglicerides | Integer | 0 | Min = 33<br>Max = 598<br>Average = 124.702 |
| 19 | Platelets | Integer | 0 | Minimum = 62<br>Maximum = 721<br>Average = 257 .025 |
| 20 | Prothrombin | Real | 2 | Min = 9<br>Max = 18<br>Average = 10.732 |

## 5.2. DATA ANALYSIS

The analysis of data was performed before applying machine learning algorithms. Details as follows.

### *5.2.1. Class Imbalance*

The class imbalance of Indian liver patient has been shown in Figure 28: Class Imbalance for Indian liver Patient data set. It has been shown that 416 persons are liver patient and 167 are not. Similarly, in Figure 29: Class Imbalance for Cirrhosis dataset has been shown where 585 patients are facing other liver disorders only 30 patients are facing liver disorder.

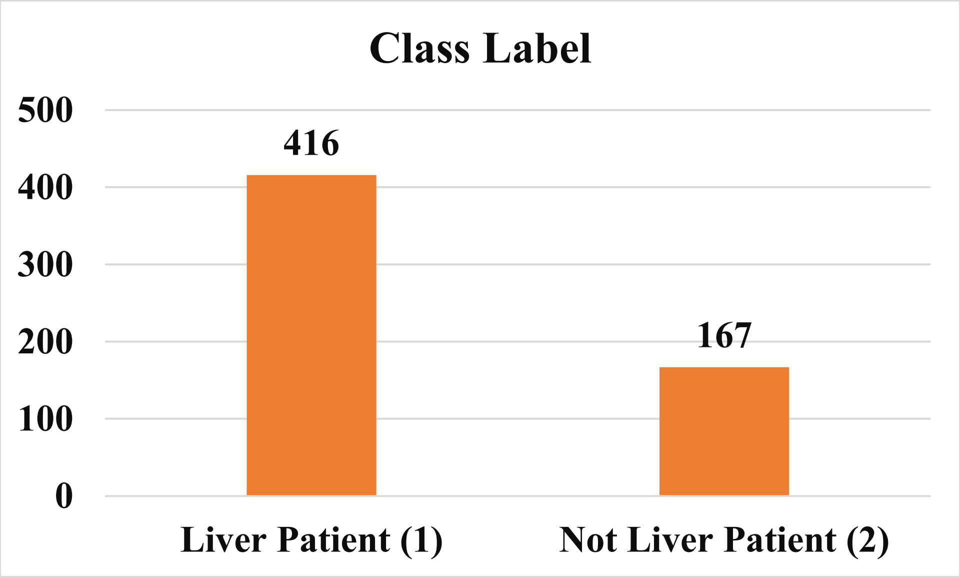


*Figure 28: Class Imbalance for Indian liver Patient data set*

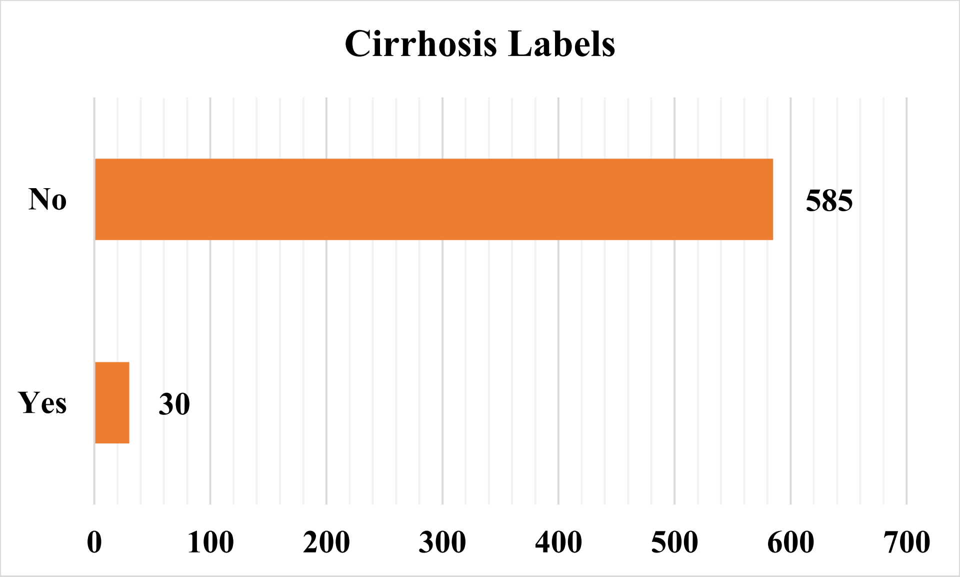


*Figure 29: Class Imbalance for Cirrhosis dataset*

### *5.2.2. Features analysis of ILPD data*

In Figure 30: Age Group frequency distribution of ILPD data it has been shown that most are adults ranging from age 21 to 72. There are 441 males and 142 females in sample dataset. In Figure 32: Boxplot of Total Protein has been shown similarly in Figure 33: Boxplot of Albumin has been shown. In Figure 34: Weight based correlation has been shown the gender, age and alkaline phosphate shown high correlation.

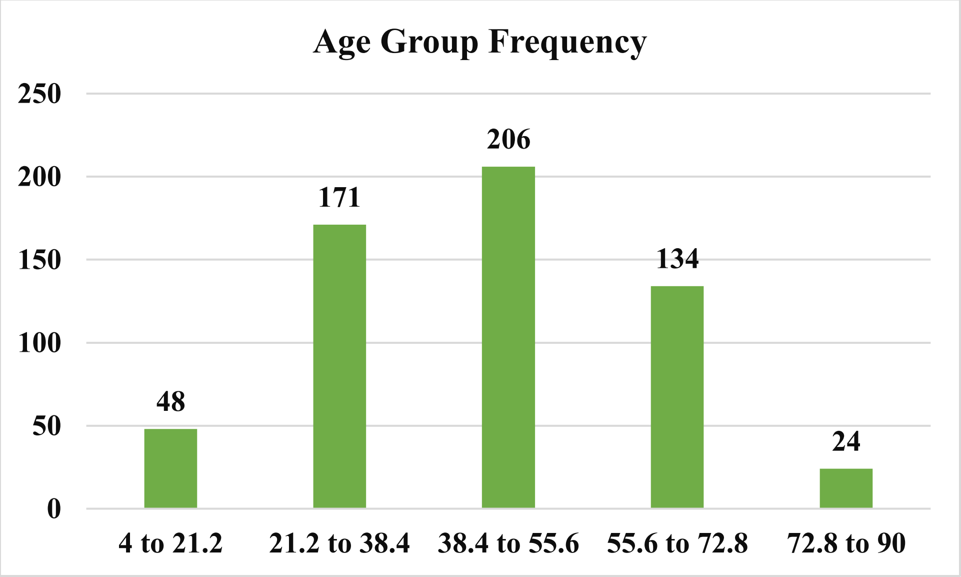


*Figure 30: Age Group frequency distribution of ILPD data*

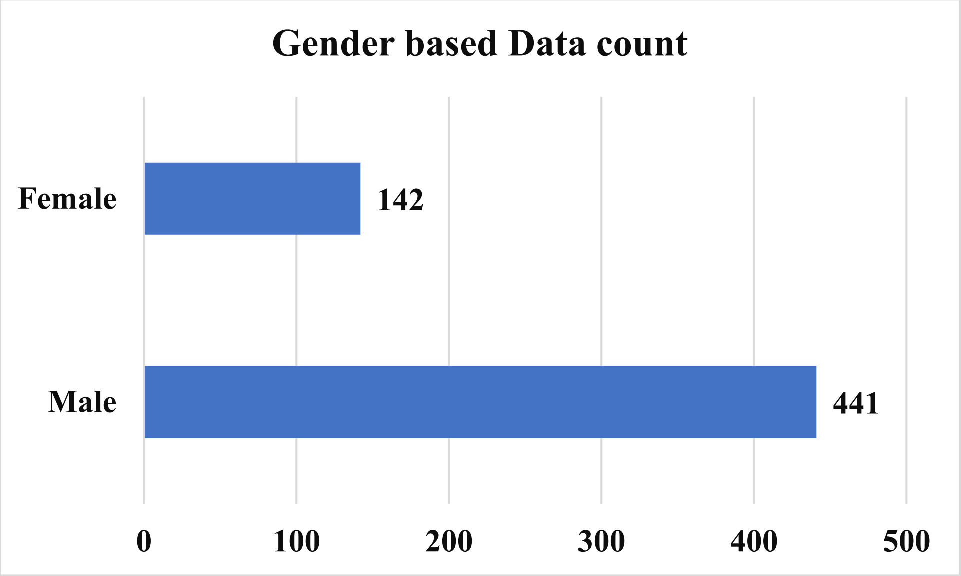


*Figure 31: Gender based data distribution*

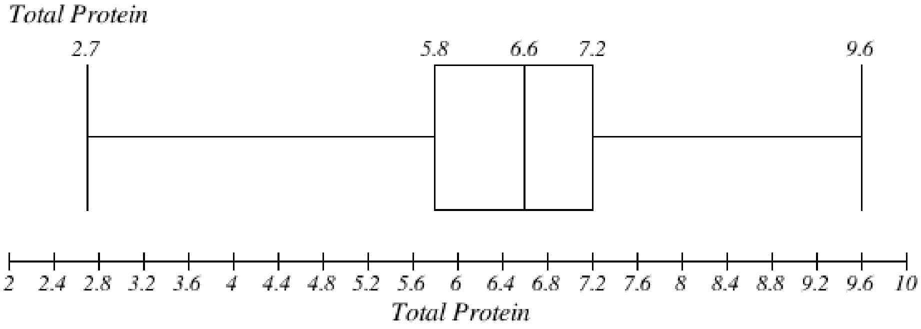


*Figure 32: Boxplot of Total Protein*

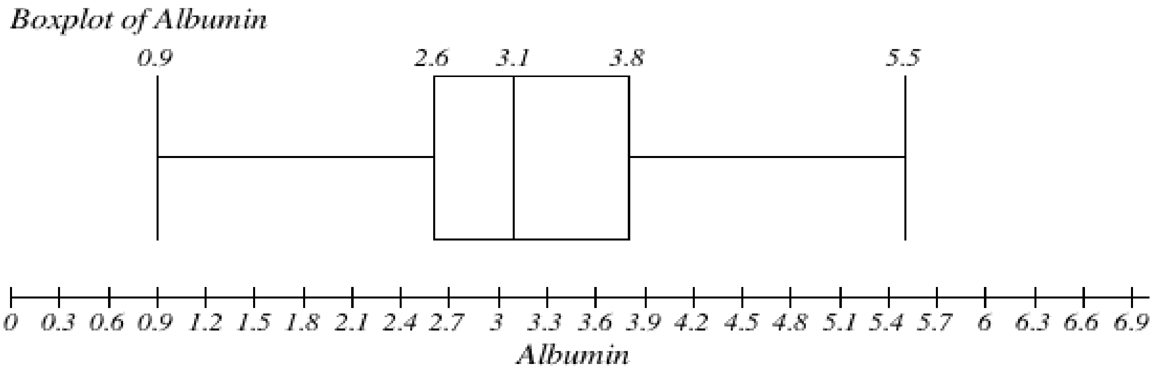


*Figure 33: Boxplot of Albumin*

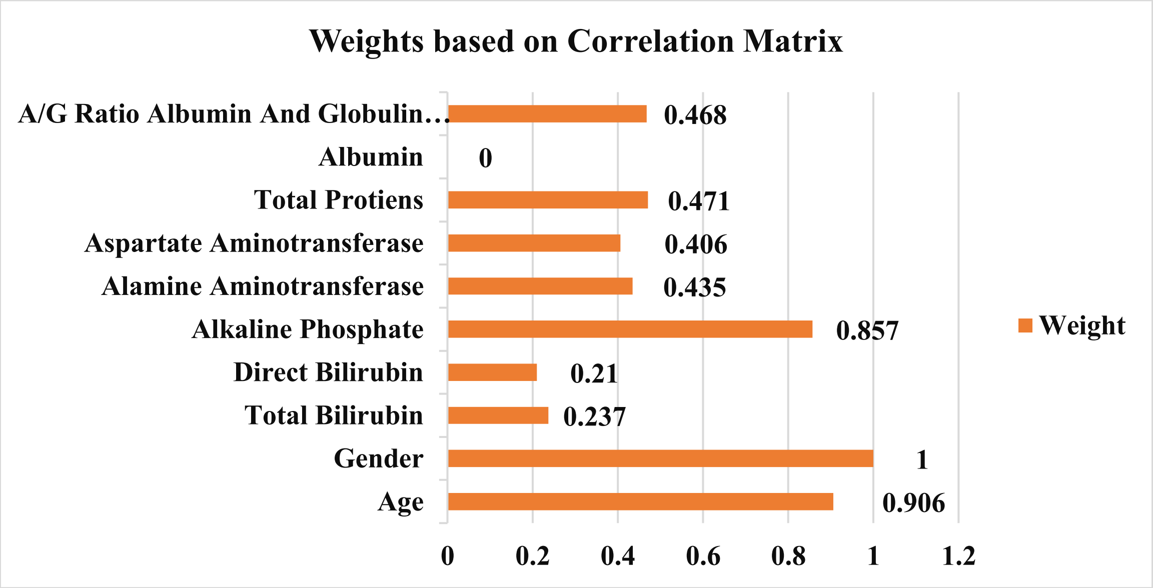
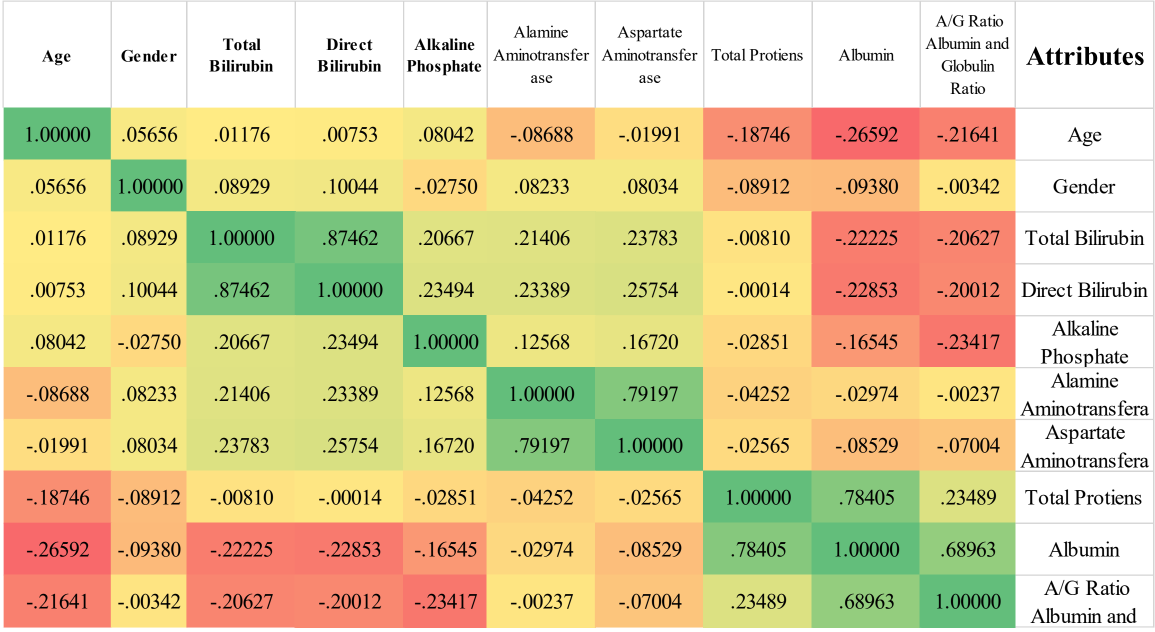


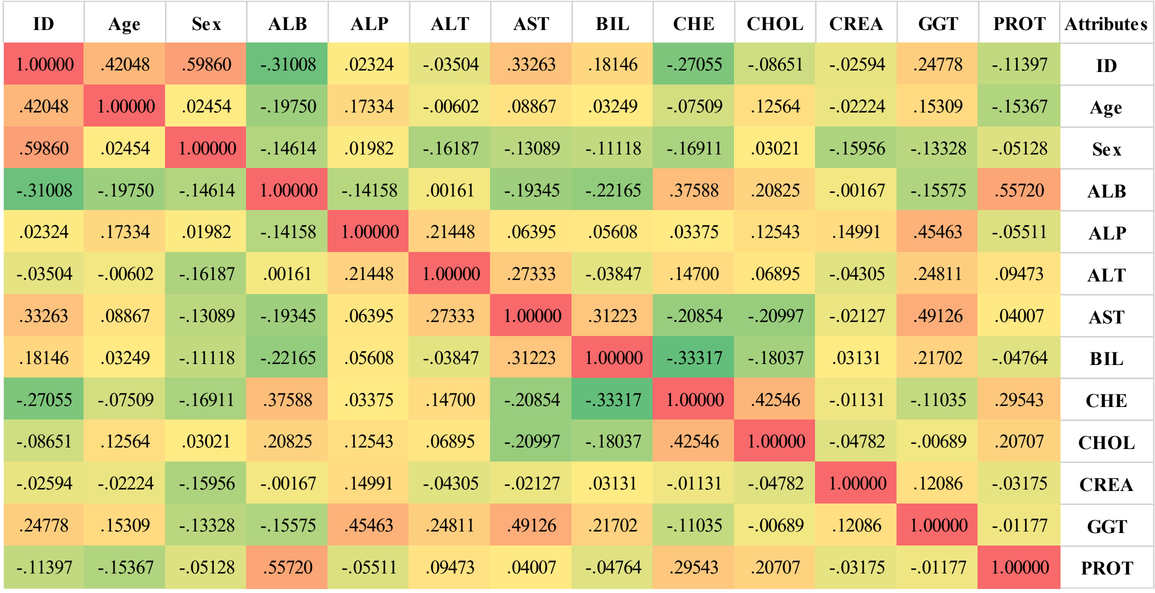
*Figure 34: Weight based correlation*

Similarly, heatmap based on correlation values shown in Figure 35: Heat Map based on correlation of ILPD dataset Similar type of analysis has been done on all data sets like Figure 36: Heat Map based on correlation of Cirrhosis data set.



| Age | Gender | Total Bilirubin | Direct Bilirubin | Alkaline Phosphate | Alamine Aminotransferase | Aspartate Aminotransferase | Total Protiens | Albumin | A/G Ratio Albumin and Globulin Ratio | Attributes |
|---|---|---|---|---|---|---|---|---|---|---|
| 1.00000 | .05656 | .01176 | .00753 | .08042 | -.08688 | -.01991 | -.18746 | -.26592 | -.21641 | Age |
| .05656 | 1.00000 | .08929 | .10044 | -.02750 | .08233 | .08034 | -.08912 | -.09380 | -.00342 | Gender |
| .01176 | .08929 | 1.00000 | .87462 | .20667 | .21406 | .23783 | -.00810 | -.22225 | -.20627 | Total Bilirubin |
| .00753 | .10044 | .87462 | 1.00000 | .23494 | .23389 | .25754 | -.00014 | -.22853 | -.20012 | Direct Bilirubin |
| .08042 | -.02750 | .20667 | .23494 | 1.00000 | .12568 | .16720 | -.02851 | -.16545 | -.23417 | Alkaline Phosphate |
| -.08688 | .08233 | .21406 | .23389 | .12568 | 1.00000 | .79197 | -.04252 | -.02974 | -.00237 | Alamine Aminotransfera |
| -.01991 | .08034 | .23783 | .25754 | .16720 | .79197 | 1.00000 | -.02565 | -.08529 | -.07004 | Aspartate Aminotransfera |
| -.18746 | -.08912 | -.00810 | -.00014 | -.02851 | -.04252 | -.02565 | 1.00000 | .78405 | .23489 | Total Protiens |
| -.26592 | -.09380 | -.22225 | -.22853 | -.16545 | -.02974 | -.08529 | .78405 | 1.00000 | .68963 | Albumin |
| -.21641 | -.00342 | -.20627 | -.20012 | -.23417 | -.00237 | -.07004 | .23489 | .68963 | 1.00000 | A/G Ratio Albumin and |

*Figure 35: Heat Map based on correlation of ILPD dataset*



| ID | Age | Sex | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT | Attributes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.00000 | .42048 | .59860 | -.31008 | .02324 | -.03504 | .33263 | .18146 | -.27055 | -.08651 | -.02594 | .24778 | -.11397 | ID |
| .42048 | 1.00000 | .02454 | -.19750 | .17334 | -.00602 | .08867 | .03249 | -.07509 | .12564 | -.02224 | .15309 | -.15367 | Age |
| .59860 | .02454 | 1.00000 | -.14614 | .01982 | -.16187 | -.13089 | -.11118 | -.16911 | .03021 | -.15956 | -.13328 | -.05128 | Sex |
| -.31008 | -.19750 | -.14614 | 1.00000 | -.14158 | .00161 | -.19345 | -.22165 | .37588 | .20825 | -.00167 | -.15575 | .55720 | ALB |
| .02324 | .17334 | .01982 | -.14158 | 1.00000 | .21448 | .06395 | .05608 | .03375 | .12543 | .14991 | .45463 | -.05511 | ALP |
| -.03504 | -.00602 | -.16187 | .00161 | .21448 | 1.00000 | .27333 | -.03847 | .14700 | .06895 | -.04305 | .24811 | .09473 | ALT |
| .33263 | .08867 | -.13089 | -.19345 | .06395 | .27333 | 1.00000 | .31223 | -.20854 | -.20997 | -.02127 | .49126 | .04007 | AST |
| .18146 | .03249 | -.11118 | -.22165 | .05608 | -.03847 | .31223 | 1.00000 | -.33317 | -.18037 | .03131 | .21702 | -.04764 | BIL |
| -.27055 | -.07509 | -.16911 | .37588 | .03375 | .14700 | -.20854 | -.33317 | 1.00000 | .42546 | -.01131 | -.11035 | .29543 | CHE |
| -.08651 | .12564 | .03021 | .20825 | .12543 | .06895 | -.20997 | -.18037 | .42546 | 1.00000 | -.04782 | -.00689 | .20707 | CHOL |
| -.02594 | -.02224 | -.15956 | -.00167 | .14991 | -.04305 | -.02127 | .03131 | -.01131 | -.04782 | 1.00000 | .12086 | -.03175 | CREA |
| .24778 | .15309 | -.13328 | -.15575 | .45463 | .24811 | .49126 | .21702 | -.11035 | -.00689 | .12086 | 1.00000 | -.01177 | GGT |
| -.11397 | -.15367 | -.05128 | .55720 | -.05511 | .09473 | .04007 | -.04764 | .29543 | .20707 | -.03175 | -.01177 | 1.00000 | PROT |

*Figure 36: Heat Map based on correlation of Cirrhosis data set*

## 5.3. RESULTS

An experiment analysis-based results has been discussed in this section categorically with their performance.

### *5.3.1. Cirrhosis data set results*

In this section performance of cirrhosis data set has been shown in Table 23: Model performance on Cirrhosis dataset.

*Table 23: Model performance on Cirrhosis dataset*

| Sr. | Algorithm | Acc | Err | Pre | F-msr | Spec |
|---|---|---|---|---|---|---|
| **1** | Decision Tree | 97.04 | 02.96 | 95.95 | 97.08 | 95.86 |
| **2** | Logistic Regression | 97.93 | 02.07 | 96.02 | 97.97 | 95.86 |
| **3** | Deep Learning | 99.11 | 00.89 | 98.26 | 99.12 | 98.22 |
| **4** | K nearest neighbors | 97.04 | 02.96 | 94.41 | 97.13 | 94.08 |
| **5** | Naïve Bayes | 95.86 | 04.14 | 93.30 | 95.98 | 92.90 |
| **6** | Random Forest | 99.70 | 00.30 | 99.41 | 99.71 | 99.41 |
| **7** | **MaLLiDD** | **99.56** | **00.44** | **99.13** | **99.56** | **99.11** |

### *5.3.2. ILPD data set results*

In this section performance of Indian liver patient data set has been shown in Table 24: Model Performance on ILPD data set.

*Table 24: Model Performance on ILPD data set*

| Sr. | Algorithm | Acc | Err | Pre | F-msr | Spec |
|---|---|---|---|---|---|---|
| **1** | Decision Tree | 69.60 | 30.4 | 62.69 | 76.10 | 42.40 |
| **2** | Logistic Regression | 74.40 | 25.60 | 69.68 | 77.14 | 62.40 |
| **3** | Deep Learning | 73.20 | 26.80 | 65.76 | 78.32 | 49.60 |
| **4** | K nearest neighbors | 75.60 | 24.40 | 70.78 | 78.14 | 64.00 |
| **5** | Naïve Bayes | 66.00 | 34.00 | 59.52 | 74.63 | 32.00 |
| **6** | Random Forest | 71.60 | 28.40 | 65.00 | 76.72 | 49.60 |
| **7** | **MaLLiDD** | **76.56** | **23.44** | **73.35** | **78.48** | **68.32** |

### *5.3.3. Cirrhosis staging dataset*

In this section performance of cirrhosis staging data set based on severity level has been shown as Table 25: Model performance on Cirrhosis staging data set

*Table 25: Model performance on Cirrhosis staging data set*

| Sr. | Algorithm | Acc | Err | Pre | F-msr | Spec |
|---|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| **1** | Decision Tree | 64.75 | 35.25 | 64.52 | 65.04 | 63.93 |
| **2** | Logistic Regression | 68.03 | 31.97 | 66.67 | 69.29 | 63.93 |
| **3** | Deep Learning | 68.03 | 31.97 | 67.19 | 68.80 | 65.57 |
| **4** | K nearest neighbors | 66.39 | 33.61 | 65.62 | 67.20 | 63.93 |
| **5** | Naïve Bayes | 68.85 | 31.15 | 66.67 | 70.77 | 62.30 |
| **6** | **MaLLiDD** | **76.11** | **23.89** | **74.41** | **77.01** | **71.75** |

## 5.4. TREE PRODUCED IN RANDOM FOREST

As it has already been discussed that the random forest is the part of ensemble the trees generated by model has been listed in this section.

### *5.4.1. Decision tree for cirrhosis data set*

Generated decision tree has been shown in Figure 37: Decision Tree for cirrhosis data set

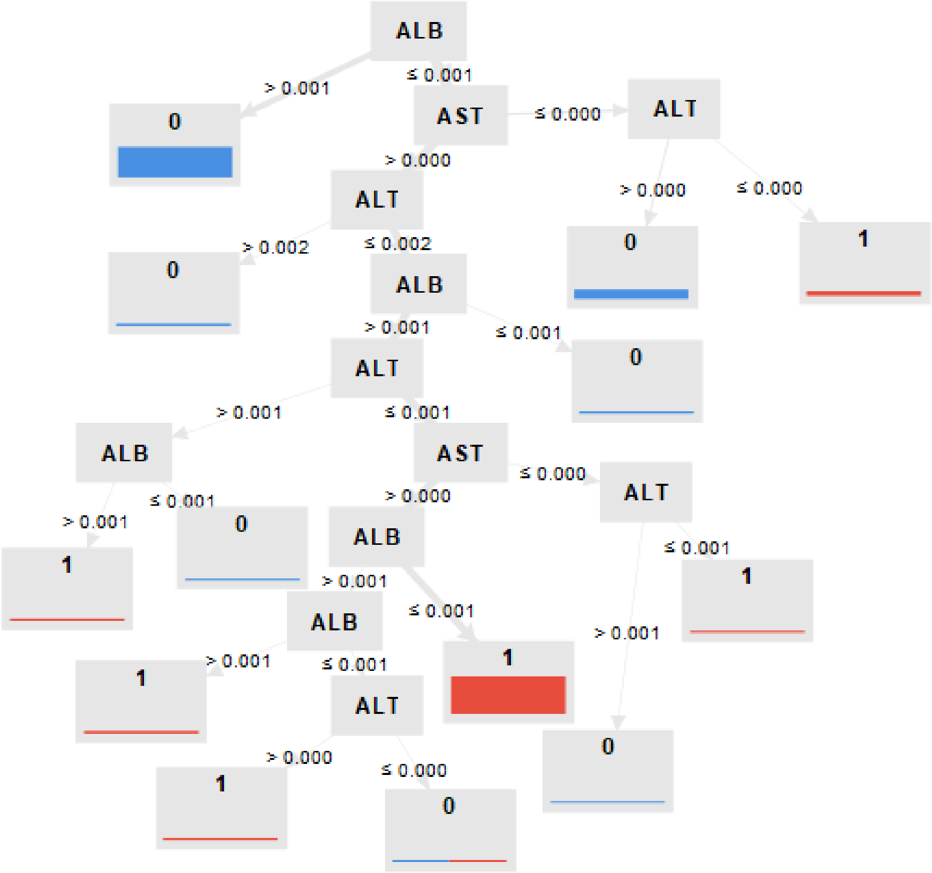


*Figure 37: Decision Tree for cirrhosis data set*

### 5.4.2. Descriptive Tree for cirrhosis data

```
ALB > 0.001: 0 {0=410, 1=0}
ALB ≤ 0.001
|   AST > 0.000
|   |   ALT > 0.002: 0 {0=13, 1=0}
|   |   ALT ≤ 0.002
|   |   |   ALB > 0.001
|   |   |   |   ALT > 0.001
|   |   |   |   |   ALB > 0.001: 1 {0=0, 1=6}
|   |   |   |   |   ALB ≤ 0.001: 0 {0=1, 1=0}
|   |   |   |   ALT ≤ 0.001
|   |   |   |   |   AST > 0.000
|   |   |   |   |   |   ALB > 0.001
|   |   |   |   |   |   |   ALB > 0.001: 1 {0=0, 1=19}
|   |   |   |   |   |   |   ALB ≤ 0.001
|   |   |   |   |   |   |   |   ALT > 0.000: 1 {0=0, 1=8}
|   |   |   |   |   |   |   |   ALT ≤ 0.000: 0 {0=1, 1=1}
|   |   |   |   |   |   ALB ≤ 0.001: 1 {0=0, 1=503}
|   |   |   |   |   AST ≤ 0.000
|   |   |   |   |   |   ALT > 0.001: 0 {0=1, 1=0}
|   |   |   |   |   |   ALT ≤ 0.001: 1 {0=0, 1=6}
|   |   |   ALB ≤ 0.001: 0 {0=6, 1=0}
|   AST ≤ 0.000
|   |   ALT > 0.000: 0 {0=113, 1=0}
|   |   ALT ≤ 0.000: 1 {0=0, 1=42}
```

### 5.4.3. Descriptive Tree for ILPD data set

```
Total Bilirubin > 0.003: 1 {1=65, 2=0}
Total Bilirubin ≤ 0.003
|   Alamine Aminotransferase > 0.003: 1 {1=28, 2=0}
|   Alamine Aminotransferase ≤ 0.003
|   |   Aspartate Aminotransferase > 0.005: 1 {1=5, 2=0}
|   |   Aspartate Aminotransferase ≤ 0.005
|   |   |   Alamine Aminotransferase > 0.001
|   |   |   |   Aspartate Aminotransferase > 0.000
|   |   |   |   |   Total Bilirubin > 0.003: 2 {1=0, 2=1}
|   |   |   |   |   Total Bilirubin ≤ 0.003
|   |   |   |   |   |   Age > 0.000
|   |   |   |   |   |   |   Alamine Aminotransferase > 0.003: 2 {1=0,
2=1}
|   |   |   |   |   |   |   Alamine Aminotransferase ≤ 0.003
|   |   |   |   |   |   |   |   Total Bilirubin > 0.001: 1 {1=30, 2=0}
|   |   |   |   |   |   |   |   Total Bilirubin ≤ 0.001: 1 {1=31, 2=6}
|   |   |   |   |   |   Age ≤ 0.000: 2 {1=0, 2=1}
|   |   |   |   Aspartate Aminotransferase ≤ 0.000: 2 {1=0, 2=3}
|   |   |   Alamine Aminotransferase ≤ 0.001
|   |   |   |   Aspartate Aminotransferase > 0.002: 1 {1=3, 2=0}
|   |   |   |   Aspartate Aminotransferase ≤ 0.002
|   |   |   |   |   Alamine Aminotransferase > 0.000
|   |   |   |   |   |   Aspartate Aminotransferase > 0.000
|   |   |   |   |   |   |   Age > 0.002: 1 {1=2, 2=0}
|   |   |   |   |   |   |   Age ≤ 0.002
|   |   |   |   |   |   |   |   Aspartate Aminotransferase > 0.000: 2
{1=236, 2=404}
|   |   |   |   |   |   |   |   Aspartate Aminotransferase ≤ 0.000: 2
{1=0, 2=3}
|   |   |   |   |   |   Aspartate Aminotransferase ≤ 0.000: 1 {1=1,
2=0}
|   |   |   |   |   Alamine Aminotransferase ≤ 0.000: 2 {1=0, 2=12}
```

### 5.4.4. Decision Tree for ILPD data set

In this section the decision tree for Indian liver patient data set has been shown as Figure 38: Decision Tree for ILPD data set .

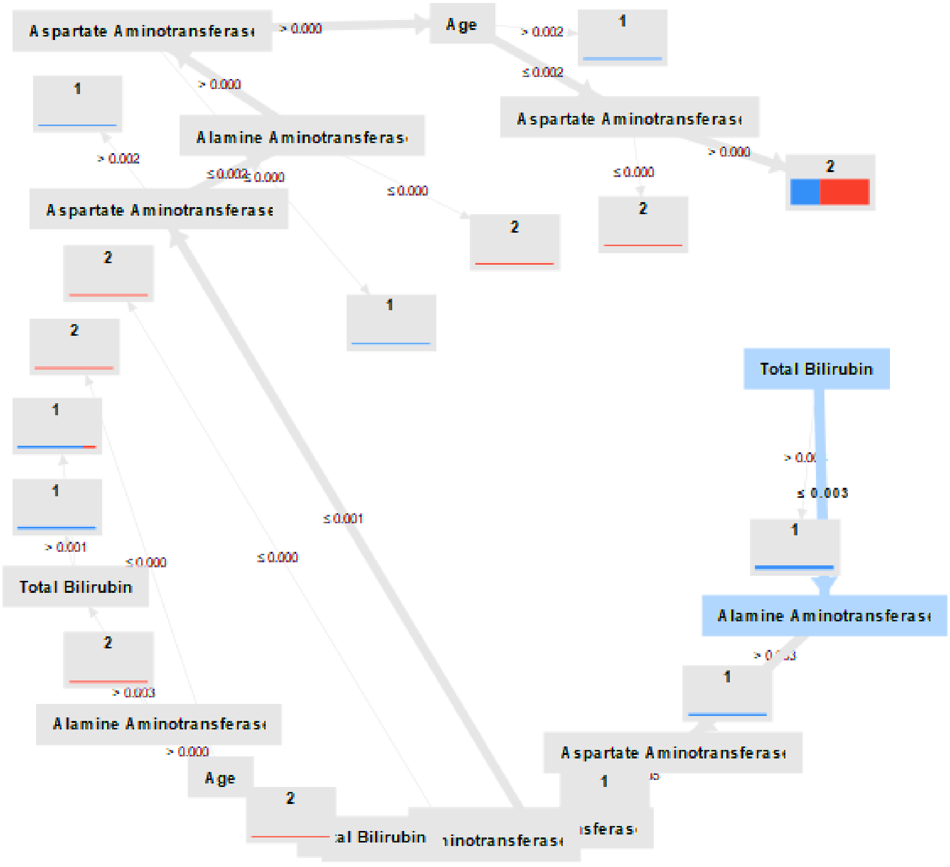


*Figure 38: Decision Tree for ILPD data set*

## 5.5. PERFORMANCE VECTORS

In this section the performance vectors have been given. The vectors have been exported from the machine learning models

### *5.5.1. Performance of cirrhosis without feature selection and sampling*

Without Feature Selection and Sampling:

Note: Without Sampling Models are Over Fitting

| Sr. | Algorithm | Acc | Err | Pre | F-msr | Spec |
|---|---|---|---|---|---|---|
| **1** | Decision Tree | 100.0 | 00.00 | 100.0 | 100.0 | 100.0 |
| **2** | Logistic Regression | 96.74 | 03.26 | 66.67 | 66.67 | 98.29 |
| **3** | Deep Learning (ANN) | 97.83 | 02.17 | 100.0 | 71.43 | 100.0 |
| **4** | K nearest neighbors | 96.20 | 03.80 | 100.0 | 36.36 | 100.0 |
| **5** | Naïve Bayes | 98.37 | 01.63 | 100.0 | 80.00 | 100.0 |
| **6** | Random Forest | 100.0 | 00.00 | 100.0 | 100.0 | 100.0 |
| **7** | MaLLiDD | 97.28 | 02.72 | 100.0 | 61.54 | 100.0 |

```
Confusion Matrix without feature selection:
   1. Decision Tree
      True:    0      1
      0:       175    0
      1:       0      9
   2. Logistic Regression
      True:    0      1
      0:       172    3
      1:       3      6
   3. Deep Learning
      True:    0      1
      0:       175    4
      1:       0      5
   4. K Nearest Neighbors
      True:    0      1
      0:       175    7
      1:       0      2
   5. Naïve Bayes
      True:    0      1
      0:       175    3
      1:       0      6
   6. Random Forest
      True:    0      1
      0:       175    0
      1:       0      9
   7. MaLLiDD without feature selection and sampling
      True:    0      1
      0:       175    5
      1:       0      4
```

### *5.5.2. Performance of Cirrhosis with feature selection and sampling*

```
The selected features are 5 in count and listed below
      ALB, ALT, AST, BIL, PROT
```

| Sr. | Algorithm | Acc | Err | Pre | F-msr | Spec | Features |
|---|---|---|---|---|---|---|---|
| **1** | Decision Tree | 97.04 | 02.96 | 95.95 | 97.08 | 95.86 | 3 |
| **2** | Logistic Regression | 97.93 | 02.07 | 96.02 | 97.97 | 95.86 | 3 |
| **3** | Deep Learning | 99.11 | 00.89 | 98.26 | 99.12 | 98.22 | 5 |
| **4** | K nearest neighbors | 97.04 | 02.96 | 94.41 | 97.13 | 94.08 | 3 |
| **5** | Naïve Bayes | 95.86 | 04.14 | 93.30 | 95.98 | 92.90 | 3 |
| **6** | Random Forest | 99.41 | 00.59 | 98.83 | 99.41 | 98.82 | 4 |
| **7** | **MaLLiDD** | **99.56** | **00.44** | **99.13** | **99.56** | **99.11** | **5** |

```
Performance Vector:
accuracy: 99.56% +/- 0.47% (micro average: 99.56%)
Confusion Matrix:
True:   0       1
0:      560     0
1:      5       565
classification error: 0.44% +/- 0.47% (micro average: 0.44%)
Confusion Matrix:
True:   0       1
0:      560     0
1:      5       565
precision: 99.13% +/- 0.91% (micro average: 99.12%) (positive class: 1)
Confusion Matrix:
True:   0       1
0:      560     0
1:      5       565
recall: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: 1)
Confusion Matrix:
True:   0       1
0:      560     0
1:      5       565
F measure: 99.56% +/- 0.46% (micro average: 99.56%) (positive class: 1)
Confusion Matrix:
True:   0       1
0:      560     0
1:      5       565
specificity: 99.11% +/- 0.94% (micro average: 99.12%) (positive class:
1)
Confusion Matrix:
True:   0       1
0:      560     0
1:      5       565
```

### *5.5.3. Performance of ILPD without feature selection and sampling*

| **Sr.** | **Algorithm** | **Acc** | **Err** | **Pre** | **Rec** | **F-msr** | **Sen** | **Spec** |
|---|---|---|---|---|---|---|---|---|
| **1** | Decision Tree | 71.84 | 28.16 | 60.00 | 06.00 | 10.91 | 06.00 | 98.39 |
| **2** | Logistic Regression | 70.69 | 29.31 | 48.48 | 32.00 | 38.55 | 32.00 | 86.29 |
| **3** | Deep Learning | 64.37 | 35.63 | 43.18 | 76.00 | 55.07 | 76.00 | 59.68 |
| **4** | K nearest neighbors | 70.69 | 29.31 | 48.48 | 32.00 | 38.55 | 32.00 | 86.29 |
| **5** | Naïve Bayes | 52.30 | 47.70 | 37.21 | 96.00 | 53.63 | 96.00 | 34.68 |
| **6** | Random Forest | 71.84 | 28.16 | 57.14 | 08.00 | 14.04 | 08.00 | 97.58 |
| **7** | **MaLLiDD** | **72.37** | **27.63** | **51.55** | **41.36** | **45.29** | **41.36** | **84.80** |

```
1. Decision Tree
   Confusion Matrix:
   True:    1      2
   1:       122    47
   2:       2      3
2. Logistic Regression
   Confusion Matrix:
   True:    1      2
   1:       107    34
   2:       17     16
3. Deep Learning
   Confusion Matrix:
   True:    1      2
   1:       66     11
   2:       58     39
4. K Nearest Neighbors
   Confusion Matrix:
   True:    1      2
   1:       107    34
   2:       17     16
5. Naïve Byes
   Confusion Matrix:
   True:    1      2
   1:       43     2
   2:       81     48
6. Random Forest
   Confusion Matrix:
   True:    1      2
   1:       121    46
   2:       3      4
7. MaLLiDD
   Confusion Matrix
   True:    1      2
   1:       351    97
   2:       63     68
```

### *5.5.4. Performance of ILPD with feature selection and sampling*

| Sr. | Algorithm | Acc | Err | Pre | F-msr | Spec |
|---|---|---|---|---|---|---|
| **1** | Decision Tree | 69.60 | 30.4 | 62.69 | 76.10 | 42.40 |
| **2** | Logistic Regression | 74.40 | 25.60 | 69.68 | 77.14 | 62.40 |
| **3** | Deep Learning | 73.20 | 26.80 | 65.76 | 78.32 | 49.60 |
| **4** | K nearest neighbors | 75.60 | 24.40 | 70.78 | 78.14 | 64.00 |
| **5** | Naïve Bayes | 66.00 | 34.00 | 59.52 | 74.63 | 32.00 |
| **6** | Random Forest | 71.60 | 28.40 | 65.00 | 76.72 | 49.60 |
| **7** | **MaLLiDD** | **76.56** | **23.44** | **73.35** | **78.48** | **68.32** |

```
Performance Vector:
accuracy: 76.56% +/- 7.67% (micro average: 76.56%)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
classification error: 23.44% +/- 7.67% (micro average: 23.44%)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
precision: 73.35% +/- 8.23% (micro average: 72.78%) (positive class: 2)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
recall: 84.83% +/- 5.86% (micro average: 84.86%) (positive class: 2)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
F measure: 78.48% +/- 6.40% (micro average: 78.36%) (positive class: 2)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
sensitivity: 84.83% +/- 5.86% (micro average: 84.86%) (positive class:
2)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
specificity: 68.32% +/- 12.43% (micro average: 68.27%) (positive class:
2)
Confusion Matrix:
True:   1       2
1:      284     63
2:      132     353
```

CHAPTER 6

# CONCLUSION AND FUTURE WORK

## 6.1. CONCLUSION

It has been concluded that the proposed MaLLiDD framework is outperforming on cirrhosis diagnostic dataset where it is showing 99.56 %. Although the model is able to run on different other data set such as Indian liver patient data set and cirrhosis data set as well. The framework is based on an ensemble of random forest, ANN and KNN algorithm. Two feature selection techniques have been used manual one based on feature importance extracted with the help of SLR, and secondly the other feature selection technique is optimized feature selection brute force. This research is categorically divided in to two parts first the SLR and the second is proposing the framework that can be capable of classifying other three datasets.

## 6.2. CONTRIBUTION

There are three major contributions of this research

- ✓ A systematic literature review for identification of outperforming classifiers and feature selection technique, Important features, effective Data Imbalance technique, and available Data sets.
- ✓ A MaLLiDD framework that has been tested on three datasets, showing effective results. Framework shows 99.56 percent of accuracy on cirrhosis data set.

## 6.3. FUTURE WORK

In this research three datasets have been tested in future more datasets can be included in test phase to evaluate the framework. An application can be developed based on this model to use it in hospitals. The framework can be enhanced to make it outperforming for all datasets.

# References

[1] N. Li *et al.*, “Machine Learning Assessment for Severity of Liver Fibrosis for Chronic HBV Based on Physical Layer With Serum Markers,” *IEEE Access*, vol. 7, pp. 124351–124365, 2019, doi: 10.1109/ACCESS.2019.2923688.

[2] R. A. Khan, Y. Luo, and F.-X. Wu, “Machine learning based liver disease diagnosis: A systematic review,” *Neurocomputing*, vol. 468, pp. 492–509, 2022, doi: https://doi.org/10.1016/j.neucom.2021.08.138.

[3] H. A, M. L, J. AM, M. M, A. GP, and V. S, “Community-Based Assessment and Treatment of Hepatitis C Virus-Related Liver Disease, Injecting Drug and Alcohol Use Amongst People Who Are Homeless: A Systematic Review and Meta-Analysis,” *International Journal of Drug Policy*, vol. 96, p. 103342, 2021, doi: https://doi.org/10.1016/j.drugpo.2021.103342.

[4] B. Moradi Kelardeh, S. Rahmati-Ahmadabad, P. Farzanegi, M. Helalizadeh, and M.-A. Azarbayjani, “Effects of non-linear resistance training and curcumin supplementation on the liver biochemical markers levels and structure in older women with non-alcoholic fatty liver disease,” *Journal of Bodywork and Movement Therapies*, vol. 24, no. 3, pp. 154–160, 2020, doi: https://doi.org/10.1016/j.jbmt.2020.02.021.

[5] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, “Burden of liver diseases in the world,” *Journal of hepatology*, vol. 70, no. 1, pp. 151–171, 2019.

[6] S. G. Sepanlou *et al.*, “The global, regional, and national burden of cirrhosis by cause in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017,” *The Lancet gastroenterology & hepatology*, vol. 5, no. 3, pp. 245–266, 2020.

[7] D. Goldberg *et al.*, “Changes in the prevalence of hepatitis C virus infection, nonalcoholic steatohepatitis, and alcoholic liver disease among patients with cirrhosis or liver failure on the waitlist for liver transplantation,” *Gastroenterology*, vol. 152, no. 5, pp. 1090–1099, 2017.

[8] H. Ma, C. Xu, Z. Shen, C. Yu, and Y. Li, “Application of machine learning techniques for clinical predictive modeling: a cross-sectional study on nonalcoholic fatty liver disease in China,” *BioMed research international*, vol. 2018, 2018.

[9] A. Spann *et al.*, “Applying machine learning in liver disease and transplantation: a comprehensive review,” *Hepatology*, vol. 71, no. 3, pp. 1093–1105, 2020.

[10] Y.-X. Liu *et al.*, “Comparison and development of advanced machine learning tools to predict nonalcoholic fatty liver disease: an extended study,” *Hepatobiliary & Pancreatic Diseases International*, vol. 20, no. 5, pp. 409–415, 2021.

[11] L. Yu, L. Jiang, D. Wang, and L. Zhang, “Attribute value weighted average of one-dependence estimators,” *Entropy*, vol. 19, no. 9, p. 501, 2017.

[12] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi, and J. Akinjobi, “Supervised machine learning algorithms: classification and comparison,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.

[13] Y. Khourdifi and M. Bahaj, “Applying best machine learning algorithms for breast cancer prediction and classification,” in *2018 International conference on electronics, control, optimization and computer science (ICECOCS)*, 2018, pp. 1–5.

[14] Z. Wang, Q. Liu, L. Wang, R. G. Gilbert, and M. A. Sullivan, “Optimization of liver glycogen extraction when considering the fine molecular structure,” *Carbohydrate Polymers*, vol. 261, p. 117887, 2021.

[15] D. Balci and E. O. Kirimker, “Hepatic vein in living donor liver transplantation,” *Hepatobiliary & Pancreatic Diseases International*, 2020.

[16] L. Feng *et al.*, “Bile acid metabolism dysregulation associates with cancer cachexia: roles of liver and gut microbiome,” *Journal of Cachexia, Sarcopenia and Muscle*, vol. 12, no. 6, pp. 1553–1569, 2021.

[17] B. Kitchenham and S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” 2007.

[18] W. Afzal, R. Torkar, and R. Feldt, “A systematic review of search-based testing for non-functional system properties,” *Information and Software Technology*, vol. 51, no. 6, pp. 957–976, 2009, doi: https://doi.org/10.1016/j.infsof.2008.12.005.

[19] L. A. Auxilia, “Accuracy prediction using machine learning techniques for indian patient liver disease,” in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 45–50.

[20] J. Singh, S. Bagga, and R. Kaur, “Software-based prediction of liver disease with feature selection and classification techniques,” *Procedia Computer Science*, vol. 167, pp. 1970–1980, 2020.

[21] R. Choudhary, T. Gopalakrishnan, D. Ruby, A. Gayathri, V. S. Murthy, and R. Shekhar, “An Efficient Model for Predicting Liver Disease Using Machine Learning,” *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 443–457, 2021.

[22] H. Hartatik, M. B. Tamam, and A. Setyanto, “Prediction for Diagnosing Liver Disease in Patients using KNN and Na\"\ive Bayes Algorithms,” in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2020, pp. 1–5.

[23] G. Shobana and K. Umamaheswari, “Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling,” in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1223–1229.

[24] L. Syafa’ah, Z. Zulfatman, I. Pakaya, and M. Lestandy, “Comparison of Machine Learning Classification Methods in Hepatitis C Virus,” *Jurnal Online Informatika*, vol. 6, no. 1, pp. 73–78, 2021.

[25] F. B. Mostafa and E. Hasan, “Machine Learning Approaches for Binary Classification to Discover Liver Diseases using Clinical Data,” *medRxiv*, 2021.

[26] S. Gupta and G. Sikka, “Explaining HCV prediction using LIME model,” in *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, 2021, pp. 227–231.

[27] D. Chicco and G. Jurman, “An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis,” *IEEE Access*, vol. 9, pp. 24485–24498, 2021.

[28] H. Ayeldeen, O. Shaker, G. Ayeldeen, and K. M. Anwar, “Prediction of liver fibrosis stages by machine learning model: A decision tree approach,” in *2015 Third World Conference on Complex Systems (WCCS)*, 2015, pp. 1–6.

[29] S. Hashem *et al.*, “Machine learning prediction models for diagnosing hepatocellular carcinoma with HCV-related chronic liver disease,” *Computer methods and programs in biomedicine*, vol. 196, p. 105551, 2020.

[30] H. Ma, C. Xu, Z. Shen, C. Yu, and Y. Li, “Application of machine learning techniques for clinical predictive modeling: a cross-sectional study on nonalcoholic fatty liver disease in China,” *BioMed research international*, vol. 2018, 2018.

[31] R. Deo and S. Panigrahi, “Prediction of hepatic steatosis (fatty liver) using machine learning,” in *Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics*, 2019, pp. 8–12.

[32] S. Agarwal *et al.*, “Development of a machine learning model to predict bleed in esophageal varices in compensated advanced chronic liver disease: A proof of concept.,” *Journal of Gastroenterology and Hepatology*, 2021.

[33] E. H. Abdelaziz, S. M. Kamal, K. El-Bhanasy, and R. Ismail, “The application of data mining techniques and feature selection methods in the risk classification of

Egyptian liver cancer patients using clinical and genetic data," in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, 2019, pp. 200–205.

[34] S. Hashem *et al.*, "Comparison of machine learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 3, pp. 861–868, 2017.

[35] K. Idris and S. Bhoite, "Applications of machine learning for prediction of liver disease," *Int. J. Comput. Appl. Technol. Res*, vol. 8, no. 9, pp. 394–396, 2019.

[36] S. Sontakke, J. Lohokare, and R. Dani, "Diagnosis of liver diseases using machine learning," in *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*, 2017, pp. 129–133.

[37] K. Babu, "A Critical Study on Cluster Analysis Methods to Extract Liver Disease Patterns in Indian Liver Patient Data," *International Journal of Computational Intelligence Research*, vol. 13, no. 10, pp. 2379–2390, 2017.

[38] M. S. Azam, A. Rahman, S. M. H. S. Iqbal, and M. T. Ahmed, "Prediction of liver diseases by using few machine learning based approaches," *Aust. J. Eng. Innov. Technol*, vol. 2, no. 5, pp. 85–90, 2020.

[39] M. Pasha and M. Fatima, "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection.," *J. Softw.*, vol. 12, no. 12, pp. 923–933, 2017.

[40] D. A. Ş. Bihter, "A Comparative Study on the Performance of Classification Algorithms for Effective Diagnosis of Liver Diseases," *Sakarya University Journal of Computer and Information Sciences*, vol. 3, no. 3, pp. 366–375, 2020.

[41] S. Gupta, G. Karanth, N. Pentapati, and V. R. B. Prasad, "A Web Based Framework for Liver Disease Diagnosis using Combined Machine Learning Models," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 421–428.

[42] C. Geetha and A. R. Arunachalam, "Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–4.

[43] M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative analysis of machine learning techniques for indian liver disease patients," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 778–782.

[44] S. Ambesange, R. Nadagoudar, R. Uppin, V. Patil, S. Patil, and S. Patil, "Liver Diseases Prediction using KNN with Hyper Parameter Tuning Techniques," in *2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC)*, 2020, pp. 1–6.

[45] S. H. Adil, M. Ebrahim, K. Raza, S. S. A. Ali, and M. A. Hashmani, "Liver patient classification using logistic regression," in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018, pp. 1–5.

[46] A. Sokoliuk, G. Kondratenko, I. Sidenko, Y. Kondratenko, A. Khomchenko, and I. Atamanyuk, "Machine Learning Algorithms for Binary Classification of Liver Disease," in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, 2020, pp. 417–421.

[47] A. S. Singh, M. Irfan, A. Chowdhury, and others, "Prediction of liver disease using classification algorithms," in *2018 4th international conference on computing communication and automation (ICCCA)*, 2018, pp. 1–3.

[48] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework," *Journal of biomedical informatics*, vol. 59, pp. 185–200, 2016.

[49] S. Ambesange, A. Vijayalaxmi, R. Uppin, S. Patil, and V. Patil, "Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques," in *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, 2020, pp. 98–102.

[50] K. Ahammed, M. S. Satu, M. I. Khan, and M. Whaiduzzaman, "Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying Machine Learning Methods," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 1371–1374.

[51] C.-C. Wu *et al.*, "Prediction of fatty liver disease using machine learning algorithms," *Computer methods and programs in biomedicine*, vol. 170, pp. 23–29, 2019.

[52] X. Pei, Q. Deng, Z. Liu, X. Yan, and W. Sun, "Machine Learning Algorithms for Predicting Fatty Liver Disease," *Annals of Nutrition and Metabolism*, vol. 77, no. 1, pp. 38–45, 2021.

[53] M. Chen and X. Zhao, "Fatty liver disease prediction based on multi-layer random forest model," in *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence*, 2018, pp. 364–368.

[54] V. V. P. Wibowo, Z. Rustam, S. Hartini, Q. S. Setiawan, and J. E. Aurelia, "Comparison between Support Vector Machine and Random Forest for Hepatocellular Carcinoma (HCC) Classification," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, 2020, pp. 618–622.

[55] Z. Cao, "Identification of the Association between Hepatitis B Virus and Liver Cancer using Machine Learning Approaches based on Amino Acid," in *Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics*, 2020, pp. 56–63.

[56] H. Che, L. G. Brown, D. J. Foran, J. L. Nosher, and I. Hacihaliloglu, “Liver disease classification from ultrasound using multi-scale CNN,” *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–12, 2021.

[57] R. Naseem *et al.*, “Performance Assessment of Classification Algorithms on Early Detection of Liver Syndrome,” *Journal of Healthcare Engineering*, vol. 2020, 2020.

[58] M. D. S. Islam, D. Liu, K. Wang, P. Zhou, L. Yu, and D. Wu, “A case study of healthcare platform using big data analytics and machine learning,” in *Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference*, 2019, pp. 139–146.

[59] S. Kumar and S. Katyal, “Effective analysis and diagnosis of liver disorder by data mining,” in *2018 international conference on inventive research in computing applications (ICIRCA)*, 2018, pp. 1047–1051.

[60] V. J. Gogi and M. N. Vijayalakshmi, “Prognosis of liver disease: Using machine learning algorithms,” in *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, 2018, pp. 875–879.

[61] M. R. Haque, M. M. Islam, H. Iqbal, M. S. Reza, and M. K. Hasan, “Performance evaluation of random forests and artificial neural networks for the classification of liver disorder,” in *2018 international conference on computer, communication, chemical, material and electronic engineering (IC4ME2)*, 2018, pp. 1–5.

[62] Z. He, C. Chen, F. Chen, Z. Huang, X. Zhu, and H. Wang, “Research on Assisted Diagnosis Model of Cirrhosis Based on BP Neural Networks,” in *Proceedings of the 2020 International Symposium on Artificial Intelligence in Medical Sciences*, 2020, pp. 271–275.

[63] H. He and Y. Ma, “Imbalanced learning: foundations, algorithms, and applications,” 2013.