**Hepatitis A and B predictions using classification learning methods**

# Abstract

**Table of Contents**

[Abstract ii](#_Toc120976679)

[CHAPTER ONE 1](#_Toc120976680)

[Introduction 1](#_Toc120976681)

[1.1 Theoretical Background 2](#_Toc120976682)

[1.2 Problem Statement 4](#_Toc120976683)

[1.3 Aim and Objectives 4](#_Toc120976684)

[1.4 Scope and Limitation 5](#_Toc120976685)

[1.5 Significance of The Study 5](#_Toc120976686)

[CHAPTER TWO 6](#_Toc120976687)

[Literature Review 6](#_Toc120976688)

[2.1 Overview of Hepatitis 6](#_Toc120976689)

[2.1.1 Types of Hepatitis 7](#_Toc120976690)

[2.1.2 Symptoms of hepatitis 8](#_Toc120976691)

[2.2 Data Mining 9](#_Toc120976692)

[2.3 Machine Learning 13](#_Toc120976693)

[2.3.1 History of Machine Learning 13](#_Toc120976694)

[2.3.2 Machine learning tasks 13](#_Toc120976695)

[2.4 Machine Learning Algorithms 15](#_Toc120976696)

[2.4.1 Artificial Neural Network 15](#_Toc120976697)

[2.4.2 Support vector machines 16](#_Toc120976698)

[2.4.3 Random Forest Classifier 16](#_Toc120976699)

[2.4.4 Bayesian networks 16](#_Toc120976700)

[2.4.5 Genetic algorithms 17](#_Toc120976701)

[2.5 Works Related to Hepatitis Classification 17](#_Toc120976702)

[CHAPTER THREE 1](#_Toc120976703)

[Research Methodology 1](#_Toc120976704)

[3.1 Preamble 1](#_Toc120976705)

[3.1.1 Research Design Approach 1](#_Toc120976706)

[3.2 Data Collection and Preprocessing 2](#_Toc120976707)

[3.2.1 The Hepatitis (UCI Repository) Dataset 3](#_Toc120976708)

[3.2.2 Data Preprocessing 3](#_Toc120976709)

[3.3 Data Mining 4](#_Toc120976710)

[3.4 Model Design, Training, and Validation 5](#_Toc120976711)

[3.4.1 Support Vector Machine 6](#_Toc120976712)

[3.4.2 Random Forest Classifier 6](#_Toc120976713)

[3.4.3 Logistics Regression 7](#_Toc120976714)

[3.5 *Evaluation of Classification Learning Methods* 7](#_Toc120976715)

[3.5.1 Confusion matrix for the predictions 8](#_Toc120976716)

[3.6 Software Requirements 11](#_Toc120976717)

[3.7 Hardware Requirement 12](#_Toc120976718)

[CHAPTER FOUR 13](#_Toc120976719)

[Implementation And Evaluation 13](#_Toc120976720)

[4.2 Data and script availability 13](#_Toc120976721)

[4.3 Exploratory Data Analysis 14](#_Toc120976722)

[4.3.1 Formula for correlation coefficient between two variables 17](#_Toc120976723)

[4.4 Training and testing of the models 20](#_Toc120976724)

[References 23](#_Toc120976725)

# CHAPTER ONE

# Introduction

**1.0 Background of the Study**

Hepatitis B virus HBV is a partially double helix DNA virus with a unique genomic structure and replication mechanism. Due to low reverse transcriptase fidelity and rapid replication, a progeny virus genome in an infected cell may not be the same as the parental genome (Andino, & Domingo, 2015). This results in a viral ecosystem consisting of swarms of mutants, or "near-species", which are populations of genetically diverse but closely related viral variants (Caligiuri et al. events, 2016). Viral hepatitis is defined as "inflammation of the liver". The liver can become inflamed for many reasons, including drinking too much alcohol, sustaining physical trauma, developing an inflammatory response, or reacting to bacteria or viruses. A, B, C, D, and E are the five most common hepatitis viruses. The Hepatitis virus can cause fibrosis, cirrhosis, liver failure, and even liver cancer. Your liver's ability to function is impaired, making it harder for your body to filter out toxins (Spearman C.W**,**2015**).** Although hepatitis A and B both affect the liver, the two viruses are very different. Hepatitis B is a blood-borne disease, transmitted mainly through direct blood-to-blood contact with an infected person. On the other hand, hepatitis A can be transmitted by the fecal-oral route or by ingestion of contaminated food or drink.

HBV infection is still a great public health concern around the world. HBV infected about 257 million people and more than 350 million people lived with CHB (Tian et al, 2019). In general, know that Seroclarance HBsAg is a significant precious marker when treating ECHB (Lee et al, 2014). By persistent patients with HBVinfed, HBSAG's annual spontaneous serum ratio ranges from 0.45% to 2.38%, prove that Seroclarance HBsAg is a rare appearance (Wu et al, 2015). The previous search told HBsAg seroclarance, whether spontaneous or caused by drugs, better prognosis, improved liver histology, lower liver cell carcinoma (HCC) and longer longevity (Yarasuri et al, 2019). Therefore, HBsAg serum clearance is an important target for better antiviral outcomes.

Previous studies have found evidence of important viral variables and host characteristics of HBsAg serum clearance. Low blood HBsAg levels have been identified by researchers as a major contributor to HBsAg serum clearance, either alone or in combination with low serum HBV DNA loading (Chong & Hao, 2019). In terms of host attributes, one of the main determinants of HBsAg antibody seroconversion is age, trailed by gender, fatty liver, cirrhosis at baseline or during follow-up, and baseline alanine aminotransferase (ALT) levels (Chu &Liew, 2010). In recent years, machine learning algorithms in healthcare have received a lot of attention. It is effectively used as a powerful classification method for extracting useful information from high-dimensional, correlated, non-linear, and imbalanced clinical datasets, enabling accurate diagnostic and predictive decisions (Austin et. al, 2013). However, the existing HBsAg serum clearance prediction model with maximum power has not been identified.

# 1.1 Theoretical Background

Machine Learning is a branch of computer science that uses a statistical technique to enable computers to "learn" from data. (For example, consistently improve proficiency in a certain activity without being explicitly coded (Sarah et al 2017).

Arthur Samuel coined the phrase "machine learning" in 1959. Samuel (1988). Machine learning evolved from the study of pattern recognition and computational learning theory in artificial intelligence and it entails the innovative work of calculations that can gain from and anticipate information (Ron, 1998) - such calculations avoid adhering to rigorously static program guidelines by settling on information-driven expectations or choices (Dickson, 2017) by developing a Machine learning is used in a variety of computing tasks where planning and programming unambiguous algorithms with excellent execution is difficult or impossible; examples include email sifting, detection of network intruders or malicious insiders attempting an information breach, optical character recognition (OCR), learning to rank, and computer vision (He and Garcia 2019).



**Figure 1. Machine learning breakdown (towarddatascience.com)**

Computational statistics, which focuses on making predictions utilizing computers, are very nearly related (frequently duplicated) to machine learning. It is very intimately related to mathematical optimization and gives equipment, hypothesis, and application domains to the field. Machine Learning and data mining are regularly befuddled (Aswathy, Jagannath, 2017), however, the latter subfield centers on exploratory data examination and analysis and is called unsupervised machine learning. You can likewise utilize unsupervised machine learning to learn and construct basic behavioral profiles for various entities. It very well may be utilized to distinguish related oddities (Tillmann, 2015).

Machine learning is a method for creating complex models and algorithms that are appropriate for forecast in the field of data analysis. For business use, this is referred to as predictive analytics. By studying past relationships and data trends, these analytical models permit specialists, information researchers, data scientists, engineers, and experts to "produce reliable and reproducible decisions and results" and unravel "hidden insights" (Bostrom, 2011).

Clustering is a non-hierarchical approach for moving data points across clusters until similar item clusters emerge or a desired set is obtained. Assumptions regarding the data set are made by clustering algorithms. If that assumption is correct, the cluster will be good. However, satisfying all assumptions is a simple task. It may be helpful to use a combination of different clustering methods as well as altering input parameters. Association rule mining is a technique for identifying common patterns among various variables in a dataset. When compared to other classification methods, associative classification usually provides a better classification.

Developing a model for assessing and accurately predicting Hepatitis B in individuals who are at risk is the primary goal of this study. Support Vector Machine, Random Forest Classifier, and Logistics Regression models will be used in this study.

# 1.2 Problem Statement

Medical diagnosis is still considered an art, despite all of the standardization attempts, because it involves knowledge in dealing with ambiguity, which is not present in today's computing gear. Artificial intelligence is a relatively recent technique in computer science, despite the fact that the notion is not new.

Seroclearance of the hepatitis B surface antigen (HBsAg) after treatment is linked to a better prediction in patients with chronic hepatitis B. (CHB). We still have a lot of work to do to figure out how to forecast HBsAg seroclearance reliably and efficiently using available clinical data. Previous studies that developed prediction models, on the other hand, relied mostly on long-term tracking of a few key variables and standard statistical approaches, which could lead to estimates that are skewed owing to the potential for collinearity in high-dimensional medical information. While Hepatitis A is not as deadly as Hepatitis B, it is similarly significant and, due to its route of transmission, is quite easy to get. In this paper, we present an active learning-based technique (AL) for training deep neural networks for hepatitis A and B classification.

# 1.3 Aim and Objectives

The goal of this research is to build a machine-learning model for Hepatitis A and B Prediction The objectives are as follows:

* To develop a method for predicting hepatitis A and B with the aid of machine learning such as Support Vector Machine, Random Forest Classifier, and Logistics Regression.
* To demonstrate the above-mentioned inquiry using a simulation approach. And,
* To compare the research's findings to similar research in the same problem domain.

# 1.4 Scope and Limitation

This research will be limited to the examination and analysis of Hepatitis A and B risk assessment and prediction activities. Other elements of the domain are not taken into account.

# 1.5 Significance of The Study

Hepatitis A and B are becoming more common over the world. In most rich and emerging countries, the rate of occurrence has been steadily increasing. Hepatitis A and B-related mortality are also on the rise, in addition to the rising incidence. Around 7.6 million individuals died worldwide from Hepatitis A and B in 2008, with about 70% of these deaths happening in underdeveloped nations (Tian et al, 2019). Every year, more than 250,000 new cases of Hepatitis A and B are projected to be detected in Nigeria, with up to 10,000 Nigerians dying from Hepatitis A and B-related causes (FBS).

These assessments or estimates may not mirror the genuine picture since they are regularly hinged on hospital-generated data, which overlooks the many cases that are not present in medical clinics, those overseen by traditional medication specialists, and the various instances of misdiagnosis in our various neighborhood hospitals. As of 2015, the United States had roughly 230,000 new instances of hepatitis A and B each year. In the United States, this sort of Hepatitis A and B causes around 40,000 deaths every year.

As a result, providing a mechanism for assessing and forecasting this risk would be extremely beneficial to the general public, particularly Nigeria as a country. Machine learning algorithms will undoubtedly provide us with access to the available processing power of recent advances in computing technology, as well as the artificial intelligence with which to appropriately utilize this computing power for solving beneficial problems such as Hepatitis A and B.

# CHAPTER TWO

# Literature Review

# 2.1 Overview of Hepatitis

Hepatitis is an ailment wherein the liver tissue becomes swollen. Hepatitis causes yellow staining of the skin and whites of the eyes (jaundice), as well as lack of appetite, regurgitation, exhaustion, stomach, and abdominal agony, and in certain people or creatures, diarrhea (WHO, 2016). Hepatitis is declared acute assuming it clears up in a half year or chronic or persistent assuming it endures longer than that (Vos et al, 2016). Intense hepatitis can disappear all alone, progress to persistent hepatitis, or lead to acute liver failure (occasionally). Cirrhosis, liver failure, and liver cancer are generally potential results of persistent or chronic hepatitis.

Hepatitis A, B, C, D, and E are the most predominant viruses that cause hepatitis. Weighty liquor ingestion, a few medications, poisons, different contaminations, immune system diseases, and non-alcoholic steatohepatitis, (NASH) are among different reasons. Hepatitis A and E are essentially sent by corrupted food and water. Hepatitis B is essentially transmitted through sexual contact, yet it can likewise be given from mother to youngster during pregnancy or labor, and it can likewise be passed on through tainted blood (Manns et al, 2019). Hepatitis C is commonly spread by tainted blood, which can happen when intravenous medication clients share needles. Just people who have proactively been contaminated with hepatitis B can get hepatitis D.

Inoculation against hepatitis A, B, and D is conceivable. Persistent viral hepatitis can be treated with prescribed drugs. Aside from those with infections that limit their average life span, antiviral prescriptions are exhorted for all individuals with persistent hepatitis C.. Despite the fact that there is no particular treatment for NASH, it is suggested that patients participate in actual physical work, consume a nutritious eating regimen, and get more fit. Prescriptions that suppress the immune system might be utilized to treat immune system hepatitis. In both intense and constant liver failure, a liver transplant might be plausible (Lawrence, 2015).

Hepatitis A tainted around 114 million individuals worldwide in 2015, persistent hepatitis B contaminated roughly 343 million individuals, and persistent hepatitis C infected roughly 142 million individuals. NASH contaminates approximately 11 million individuals in the United States, while alcoholic hepatitis affects around 5 million. Hepatitis causes in excess of 1,000,000 fatalities every year, the vast majority of which are brought about by liver scarring or malignancy.

## 2.1.1 Types of Hepatitis

The five distinct hepatitis viruses identified by the letters A, B, C, D, and E are all responsible for liver disease, but they differ significantly.

The hepatitis A virus (HAV) is found in infected people's feces and is most frequently spread by eating or drinking contaminated food or water.

HAV can also be spread through certain forms of sex. Most people recover completely and remain immune to subsequent HAV infections, so infections tend to be mild. However, HAV infections can also be severe and life-threatening. People have contracted this virus in areas of the world with poor sanitation.

HAV can be prevented with vaccines that are both safe and effective. The hepatitis B virus (HBV) is spread by contact with infected blood, saliva, and other bodily fluids.

HBV can pass from infected mothers to infected infants at the time of birth or from family members to infected infants in the early stages of childhood. Injection drug use, contaminated injections during medical procedures, and transfusions of HBV-contaminated blood and blood products are all potential routes of transmission.

Healthcare workers who sustain accidental needle stick injuries while caring for infected HBV patients are also at risk from HBV.HBV can be prevented with vaccines that are both safe and effective.

Most of the time, the hepatitis C virus (HCV) is spread through contact with infected blood. This can happen when HCV-infected blood and blood products are transfused, contaminated injections are used during medical procedures, or injection drug use occurs. Transmission via sexual means is also possible, though it is much less common. There is no HCV vaccine.

Only people with HBV infection are infected with the hepatitis D virus (HDV). The combination of HDV and HBV infection may lead to a more severe condition with a worse outcome. HDV infection is prevented by hepatitis B vaccines.

Most people get hepatitis E virus (HEV) from eating or drinking contaminated food or water. In developing nations, HEV is a frequent cause of hepatitis outbreaks and is becoming increasingly recognized as a significant disease cause.

Although vaccines against HEV infection that are both safe and effective have been developed, they are not widely available.

## 2.1.2 Symptoms of hepatitis

A chronic form of hepatitis, such as hepatitis B and C, may not cause symptoms until damage to liver function has occurred. Acute hepatitis, on the other hand, can cause symptoms as soon as a person contracts the virus. The following are typical signs of infectious hepatitis:

* fatigue
* flu-like symptoms,
* dark urine
* pale stools
* abdominal pain
* loss of appetite
* unexplained weight loss

**Table 2.1: Causes of hepatitis**

|  |  |
| --- | --- |
| Type of hepatitis | Common route of transmission |
| hepatitis A | exposure to HAV in food or water |
| hepatitis B | contact with HBV in body fluids, such as blood, vaginal secretions, or semen |
| hepatitis C | contact with HCV in body fluids, such as blood, vaginal secretions, or semen |
| hepatitis D | contact with blood containing HDV |
| hepatitis E | exposure to HEV in food or water |

# 2.2 Data Mining

As per Manjunath et al., (2010) and Kimball R. (2000), the utilization of choice emotionally supportive mechanisms is a method that effectively utilizes information to handle, operate, and improve the assessment of educational institutions more readily. Dependent on the attribute and availability of the fundamental data, an answer like that could handle various issues by refining data from any mix of student records management systems. Mining information from an information store can be an immediately accessible and productive system for decision-makers. Manjunath et al. (2010) found that an information store is a problem engaged, unified, stable, and time-variant assortment of information used to help the executives’ decisions. Considering this perspective, Jiawei et al., (2000) depicted the engineering of an ordinary information mining system as displayed in Figure 2.1.

**Database, data stores, or other information warehouse**. This is a database, one or a variety of data sets, information stores, spreadsheets, or different kinds of data warehouses. Techniques, for example, Data cleansing and information integration can be applied to the data.

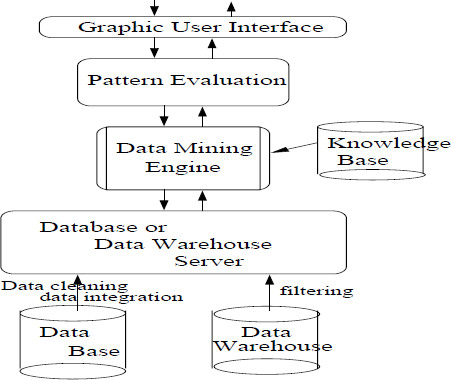
**Database or data warehouse server**. A user's data mining requirements determine whether the database server or the data warehouse server is responsible for retrieving the necessary data.

**Knowledge base**. Domain knowledge can be used to guide research or to measure whether patterns discovered are useful. Concept hierarchies can be used, for example, to organize attributes at different levels of abstraction. In addition, user confidence, which is used for judging the attractiveness of a model based on its unexpectedness, can also be calculated. Furthermore, domain knowledge can also refer to additional boundaries or thresholds of interest, as well as metadata (e.g., description of data from multiple heterogeneous sources).

**Data mining engine**. The characterization, association analysis, classification, analysis, and deviation analysis module are an important component of a data mining system.

**Pattern evaluation module**. Often, this component uses interestingness measures and works with the data mining modules to extract interesting data. It has access to the knowledge base storing the interestingness thresholds. Depending on the architecture of the data extraction method, the sample evaluation module can additionally be integrated with the extraction module. To limit mining to only interesting patterns, we recommend pushing the evaluation of pattern interestingness as deep as possible into the data mining process.

**Graphical user interface**. In this module users can specific queries or data mining tasks in order to interact with the data mining system, providing information to help them focus on their research and operations activities. This module uses intermediate data mining to determine what's worth investigating. With this component, users can also browse databases and database programs, evaluate mined patterns, and view patterns in a variety of ways.

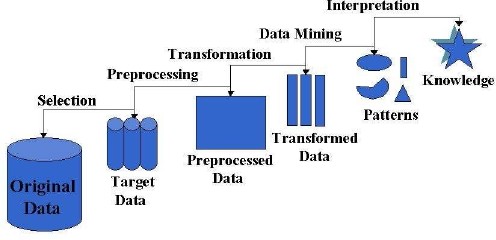


**Figure 2.1: Architecture of a typical data mining system. (Jiawei et. al., 2000)**

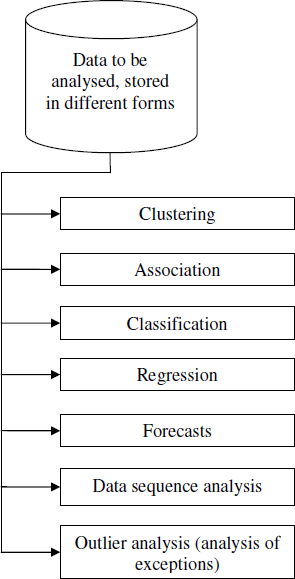
Among the procedures utilized in information, mining is database technology, statistics, machine learning, elite performance computing, pattern recognition, neural networks, data representation, data recovery, image and signal handling, and information analysis of Geographic data.

Figure 2.2 illustrates the general data mining process as described by Zarki (2003). It includes the steps below, some of which are optional depending on the situation in question:

1. Grasp the application domain: To appreciate user-requested data mining results, you must first understand the scope. To increase your chances of success, you need to be content and use all the information you have before.
2. Collect and construct the target dataset: Data mining relies on the availability of relevant data that reflects the variety, order, and underlying structure of the problem. Therefore, it is essential to collect a data set that includes all conceivable scenarios relevant to the topic under study (Jiawei et al., 2000).
3. Clean up and change the objective dataset: Raw information has numerous blunders and irregularities, like noise, exceptions or outliers, and missing qualities. Copying information records to build a non-repetitive dataset is a significant piece of this cycle (Troyanskaya et al., 2001).
4. Select elements, and limit dimensions: Even after the information has been cleaned of copies, irregularities, missing qualities, and so forth, there can in any case be noise inconsequential to the subject being studied. These noisy elements can prompt disarray in later phases of information investigation, as well as the creation of irrelative rules and connections and expanded computational expenses. Therefore, it is very reasonable to perform dimensional reduction or feature selection stage to distinguish useful attributes from irrelevant attributes (Jiawei et al., 2000). Fisher's criteria, Wilcoxon's rank sum test, principal components analysis, entropy analysis, and other statistical or heuristic techniques are commonly used to perform this stage.
5. Data mining algorithms can be used to examine data, such as association rule discovery, sequence mining, classification tree induction, clustering, etc.
6. Patterns for interpretation, evaluation, and visualization: Check output to interpret and evaluate extracted patterns, rules, and patterns. New perspectives on the situation can only be analyzed through this process of interpretation and evaluation.



**Figure 2.2: The procedure for data mining. (Zaki*et. al.,* 2003)**

****

**Figure 2.3: The most important data mining jobs for analyzing vast amounts of data (Andronie*et. al.,*2010)**

# 2.3 Machine Learning

The scientific investigation of statistical models and algorithms that are used by computer systems to achieve undertakings coherently without utilizing express guidelines, rather depending on patterns and deduction, is Machine Learning (ML). It is considered a subset of artificial intelligence. Machine learning algorithms manufacture a mathematical model of test information, known as "training data," to deliver estimates or decisions without being explicitly modified to do thusly (Koza, 1996). Machine Learning techniques are utilized in applications, for example, email separating, network intrusion discovery, and computer vision, where fostering an algorithm with explicit guidelines for task completion is not feasible. Machine Learning is nearly connected with computational statistics, which is focused on forecasting or foretelling with the aid of computers. The study of mathematical optimization helps the discipline of machine learning by giving devices, hypotheses, and application domains. Information mining is a subfield of machine learning that spotlights unaided learning for exploratory information analysis When used to tackle business issues, machine learning is additionally alluded to as prescient analytics.

## 2.3.1 History of Machine Learning

In 1959, Arthur Samuel invented the term "machine learning". In a more formal explanation of the algorithms explored in the machine learning field, Tom M. Mitchell stated:

*"A computer program is said to learn from experience E with respect to a particular class of tasks T and a particular performance measure P if the program improves its performance at those tasks P as measured by E." Mitchell, T. (1997).*

As opposed to portraying the subject in analytical terms, this meaning of machine learning tasks gives a very basic level of functional depiction. This is in accordance with Alan Turing's methodology in his paper "Computing Machinery and Intelligence," which replaces the inquiry "Can machines think?" with "Can machines do what we (as thinking creatures) can accomplish?" (Harnad 2008). The various properties that could be held by a reasoning machine, as well as the different ramifications in creating one, are introduced in Turing's proposition.

## 2.3.2 Machine learning tasks

Machine learning occupations are isolated into general classes. In controlled or supervised learning, the algorithm develops a numerical model from a lot of data that consolidates the two information sources and expected outcomes. For example, suppose the goal is to conclude whether an image contains a specific item, then, the training information for the controlled learning system will include pictures with and without the item (the input) and each image will have a name (the output) showing whether it contained the item. In various cases, the input may be confined to unequivocal reactions or just somewhat available. Semi-controlled learning procedure makes numerical models from inadequate training information, where some example inputs are unlabeled.

Supervised learning techniques incorporate classification and regression algorithms. Whenever the results are restricted to a little arrangement of values, classification techniques are applied. An incoming email will serve as the input to an email sifting classification algorithm and the folder’s name to classify the mail will serve as the output. The output of a spam email recognition algorithm would be a declaration of "spam" or "not spam," depicted by the Boolean qualities true and false. Regression algorithms obtain their name from the way that their results are continuous, meaning they can be any value within a range. Temperature, length, and item cost are generally instances of continuous qualities.

The system makes a mathematical model from an informational set that incorporates just data sources or inputs and no ideal or desired output marks for unaided learning. Unaided learning procedures are utilized to investigate the structure of information, for example, bunching or grouping of data points. Unaided learning, similar to feature learning, can find patterns in information and can sort the inputs into different categories. The process of scaling down the number of "features", or inputs, in a set of data is referred to as dimensionality reduction.

The choice inputs for which training labels are acquired are optimized by active learning algorithms. This enhancement is achieved by obtaining access to desire outputs (training labels) for a definite number of inputs in view budget. When used collaboratively, they can be given to a human client for labeling. In autonomous vehicles or while sorting out some way to play a game against a human adversary, Reinforcement learning algorithms are used, and they receive feedback as positive or negative support in an evolving environment (Bishop, 2006). Topic Modeling is a kind of Machine learning methodology wherein a computer program is given a cluster of natural language texts and asked to find other compositions or documents that pertains to related topics. Machine learning strategies may be used in density estimation problems, to acquire the imperceptible probability density function. Meta learning algorithms gain proficiency with their own inductive bias considering previous happenings. In developmental robotics, robot learning algorithms produce their own sequence of learning experience, for the most part insinuated as an instructive arrangement or educational program, to learn new limits over the long haul through autonomous investigation and social communication with people. Active learning, maturation, motor synergies, and imitation, are the means through which these robots are guided.

# 2.4 Machine Learning Algorithms

## 2.4.1 Artificial Neural Network

An artificial neural network is a combination of interrelated nodes that works almost the same way as the brain's immense collection of neurons. Each circular node portrays an artificial neuron, and the link between the input of another artificial neuron and the output of one artificial neuron is portrayed by each arrow. Artificial neural Networks (ANNs), otherwise called connectionist systems, are computational systems in view of natural neural networks detected in animal brains (Scott, 2010). The neural network is a system that permits different machine learning algorithms to team up and investigate intricate data inputs. It isn't, all by itself, an algorithm. Such frameworks "learn" to tackle undertakings by looking at samples, as often as possible with practically no task explicit instructions being programmed.

An artificial neural network (ANN) is a model that is based on a network of linked units or nodes known as "artificial neurons," which are roughly modeled after neurons in the human brain. Every link, like synapses in the human brain, may transfer data or a "signal," beginning with one artificial neuron and progressing to the next. When an artificial neuron receives a signal, it can decode it before sending it on to other artificial neurons. In ordinary ANN implementations, the signal at a link between artificial neurons is a real number, and the output of each artificial neuron is generated by some non-linear function of the sum of its inputs. "Edges" refer to the connections between artificial neurons. The weight of artificial neurons and edges is regularly modified as learning progresses. The weight affects signal strength at a link by expanding or contracting it. Artificial neurons may have a baseline that prevents signal transmission until the entire signal exceeds the baseline. Artificial neuron layers are frequently used. Various layers may make distinct adjustments to their feedback bits. Signals are guided from the first (input layer) to the final (output layer), maybe many times.

The basic goal of the ANN technique was to approach problems in the same manner that a human mind would. However, after a while, the attention shifted to certain tasks, resulting in biological abnormalities. Computer vision, voice recognition, machine interpretation, social network sorting, board, and video games, and clinical diagnostics are all examples of how artificial neural networks have been employed.

Deep learning is comprised of a few secret layers in an artificial neural network. This strategy endeavors to reenact how the human brain processes light and sound into vision and hearing. Deep learning has been utilized effectively in many aspects, including computer vision and speech recognition (Honglak, 1995).

## 2.4.2 Support vector machines

Support vector networks, or SVMs, are supervised learning techniques used for categorization and regression. A model that predicts whether a new case falls into one of the two categories given a progression of training scenarios that are labeled as having a place with one of the two classifications is assembled by an SVM training strategy (Cortes, 1995). Platt scaling, for example, is a probabilistic technique to employing SVM in a probabilistic classification environment, yet an SVM training algorithm is a non-probabilistic, paired, linear classifier. The kernel technique enables SVMs to perform both non-linear and linear characterization by verifiably interpreting their inputs into high-dimensional feature spaces.

## 2.4.3 Random Forest Classifier

Random forest, as its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction.

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well.

## 2.4.4 Bayesian networks

This is a simple straightforward Bayesian network. Whether or not the sprinklers are turned on is influenced by the rain, and both the rain and sprinklers affect how moist the grass is.

A Bayesian network, otherwise called a belief network or a directed non-cyclic graphical model, is a probabilistic graphical model that portrays a bunch of irregular factors and their contingent independence utilizing a directed acyclic chart (DAG). A Bayesian network, for instance, may be utilized to show the probability relationship among ailments and symptoms. In view of side effects, the network might be utilized to decide the probability of specific diseases being present. There are productive induction and learning algorithms. Dynamic Bayesian networks are Bayesian networks that model variable arrangements like voice signals or protein successions. Influence diagrams are Bayesian network generalizations that might depict and tackle decision issues under uncertainty.

## 2.4.5 Genetic algorithms

A Genetic algorithm (GA) is a search algorithm and heuristic instrument that recreates natural selection by creating new genotypes through techniques such as mutation and crossover with expectations of uncovering ideal solutions to a specific issue. Genetic algorithms were utilized in machine learning during the 1980s and 1990s (Zhang, 2011). Conversely, machine learning algorithms have been utilized to enhance the performance of genetic and evolutionary algorithms.

# 2.5 Works Related to Hepatitis Classification

Classification is one of the most significant and critical jobs in machine learning and data mining. A lot of research has been done on various medical datasets utilizing data mining and machine learning to categorize Hepatitis A and B. Many of them have a high level of precision in categorization.

Bandyopadhyay et al (2020) suggested an automated system that distinguishes patients with hepatitis syndromes in their work A Voting Ensemble Approach for Hepatitis Disease Detection. They subsequently developed an automated method based on a multi-phased classification technique.

Several classifiers, including the Support Vector Machine, Multi-layer Perceptron, naive Bayes, k-Nearest Neighbor, and Decision Tree, were utilized in the early phase. Ada Boost, Gradient Boost, and Random Forest were employed as phase-2 classifiers. These implemented classifiers were analyzed and compared in terms of prediction performance. They created a voting-based ensemble technique for the final phase classification, which uses the top two classifier models from the first and second phase classifications, respectively. They argue that using the proposed classifier would increase prediction performance, allowing patients with hepatitis to be properly identified.

Yarasuri, Indukuri, and Nair's paper "Prediction of Hepatitis Disease Using Machine Learning Technique" (2019). They contrasted and compared several machine learning technologies and neural networks. They used the accuracy rate and mean square error as performance measures. Machine Learning (ML) methods such as Support Vector Machines (SVM), K Nearest Neighbor (KNN), and Artificial Neural Networks (ANN) have been used to categorize and predict Hepatitis disease. They performed a brief examination into the aforementioned problem based on illness diagnostic prediction accuracy. All of the machine learning approaches were implemented and validated using the MATLAB program.

Hepatitis Disease Prediction Oishi et al. advocated using an Effective Deep Neural Network (2020). Their study objective was to build a Deep Neural Network model that could predict the illness with accuracy. Their study was compared with activation functions to assure the best performance. The study made use of a standard dataset on hepatitis infections gathered from 154 people in India. They claimed that their model's accuracy for predicting hepatitis was 92.3 percent, which they said was higher than previous research or studies using other approaches.

Some researchers have employed a multi-model method in their work Prediction of Hepatitis Disease. Some of the approaches employed include K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron, and Random Forest. They predicted hepatitis disease using a variety of data mining techniques. In addition, they proposed a viable strategy for increasing the performance of their prediction models. They dealt with missing values in their dataset by eliminating missing value observations. They were able to discover extraneous attributes by combining the info-gain feature selection approach with a ranker search. To calculate prediction accuracy on the hepatitis disease dataset, classification methods such as K-Nearest Neighbors (KNN), Naive Bayes Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest were utilized. They compared the classification models' performance by assessing accuracy, precision, recall, F1-score, and ROC. They claimed that removing observations with missing values and employing the info-gain feature selection technique in their prediction models improved their accuracy. Many found that the Random Forest model performed the best, with a classification accuracy of 92.41percent.

Gowtham et al (2021) proposed a deep learning strategy for classifying Paddy disease. Their research aimed to assist a large number of farmers, particularly those involved in paddy farming, in understanding and forecasting the illness that would affect their crops. Their study demonstrated the accuracy of using deep neural networks to classify paddy leaf disease. On their dataset, they employed pre-processing techniques such as data augmentation and the median filter to reduce overfitting and improve model performance and accuracy.

PinarYildirim created Filter-Based Feature Selection Methods for Risk Prediction in Hepatitis Disease (2015). A comparison of filter-based feature selection techniques was carried out using a well-known dataset (the hepatitis dataset), and four classification algorithms were used to test the performance of the algorithms. Following the application of feature selection methods, they claimed that Nave Bayes and Decision Table classifiers outperformed the others on the hepatitis dataset. According to them, the study indicated that feature selection techniques can improve the performance of learning algorithms. However, they claimed that no single feature selection technique based on filters is the best. Consistency Subset, Info Gain Attribute Eval, One-R Attribute Eval, and Relief Attribute Eval outperformed the others in general.

Mohamed et al. (2014) investigated the use of non-invasive liver fibrosis models to predict morbidity and mortality in chronic hepatitis C patients. They investigated Compensated CHC patients for seven years. Using Hepascore, FIB-4, APRI, and liver biopsy data, a list of non-invasive indicators was created. The Western Australian Data Linkage System was used to acquire follow-up morbidity and mortality data. The Kaplan–Meier approach was used to examine the prognostic relevance of baseline noninvasive indicators and biopsies. A total of 406 persons (64 percent male, mean age 48 11 years) were tracked for 2385 person-years, during which 22 (5.4 percent) died, 14 (3.4 percent) died from liver disease or required a liver transplant. Sixteen of the participants had hepatic decompensation (3.9 percent). Hepascore and liver biopsy (P 0.005), but not APRI or FIB-4, predicted overall and liver-related mortality and decompensation. A Hepascore greater than 0.5 was associated with an increased risk of overall mortality [HazardRatio(95 percent CI) 6.7 (2.6–17), P0.001], liver-related mortality[32.8 (4.3–250), P = 0.001], and future decompensation [11.8 (3.3–41), P 0.001], whereas a Hepascore less than 0.5 was associated with a 99 percent chance of not dying from liver-related causes over 10. With an AUROC of 0.86 (95 percent CI 0.80–0.90) and 0.87 (0.79–0.96), Hepascore was as accurate as liver biopsy in predicting liver-related mortality.

# CHAPTER THREE

# Research Methodology

# 3.1 Preamble

The methodology utilized in carrying out the project will be talked about in this part of the research study. It will be partitioned into three areas: a discourse on the proposed methodology and an overall discussion of its parts, point by point explanations of every aspect of the proposed methodology, including any pivotal algorithms, diagrams, and a discussion of the assessment strategy of the proposed methodology.

## 3.1.1 Research Design Approach

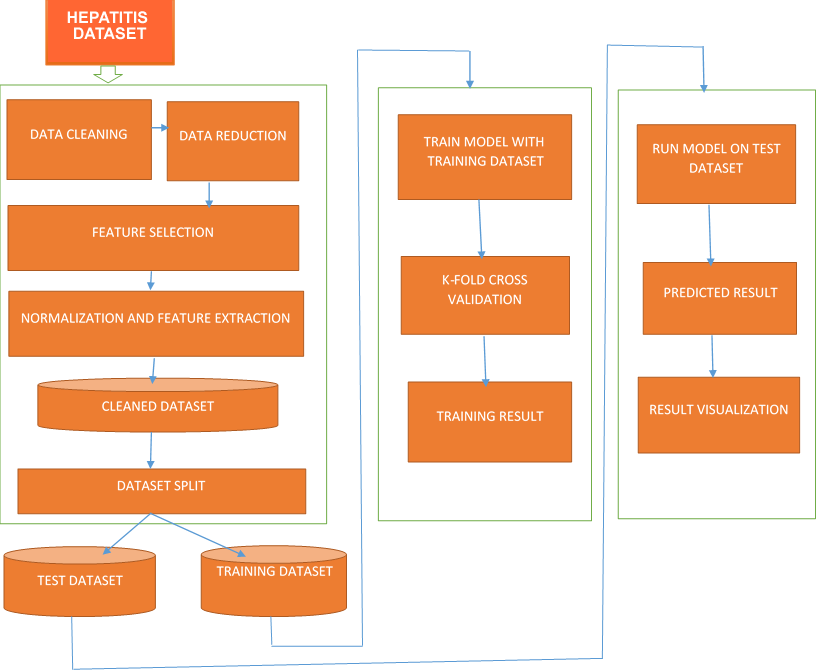
Hepatitis is one of the most common liver illnesses in the world, accounting for many new sickness cases and hepatitis-related mortality, making it a critical public health concern in today's society, according to worldwide data.

Early hepatitis identification can help improve a patient's chances of survival since it allows medical providers to deliver focused therapy based on an accurate diagnosis. To minimize persons obtaining unnecessary drugs, hepatitis A and B should be categorized more precisely. As a result, the purpose of this research is to offer a reliable diagnosis of hepatitis A and B, whether present or absent. Because of its benefits in recognizing crucial characteristics from complicated hepatitis datasets, machine learning (ML) is generally recognized as the technique of choice in hepatitis pattern categorization and forecast modeling.

A methodology is proposed that will obtain the UCI hepatitis dataset, carry out the required information preprocessing tasks on it to eliminate and data irregularities to prevent dimensionality thereby preventing model over fitting, partition the purified dataset into testing dataset and training dataset, utilizing the training dataset to train the model, and afterward validate the trained model utilizing the test dataset. The model will be built utilizing a two-dimensional convolutional neural networking conjunction with a bidirectional long-short term memory encoder. The bidirectional long-short term memory will effectively become familiar with the dataset and the two-dimensional convolutional neural network and will classify it into the suitable hepatitis A or B class.

The proposed methodology is shown in figure 3.1 below. The methodology is divided into three major parts namely:

1. Data gathering and preprocessing
2. Model plan, training, and validation
3. Hepatitis discovery and assessment technique



**Figure 3.1: Research Design**

# 3.2 Data Collection and Preprocessing

The dataset utilized for the research activity coupled with the means through which the dataset will be processed to ensure its availability for usage in our analysis, is amply discussed in this phase of the methodology. The hepatitis dataset from UCI Wisconsin is being used. The next section delves into the dataset's intricacies.

## 3.2.1 The Hepatitis (UCI Repository) Dataset

The glimpse of the data is listed as follows:

1. Class: DIE, LIVE

2. AGE: 10, 20, 30, 40, 50, 60, 70, 80

3. SEX: male, female

4. STEROID: no, yes

5.ANTIVIRALS: no, yes

6. FATIGUE: no, yes

7. MALAISE: no, yes

8. ANOREXIA: no, yes

9.LIVER BIG: no, yes

10. LIVER FIRM: no, yes

11.SPLEEN PALPABLE: no, yes

14. SPIDERS: no, yes

15. ASCITES: no, yes

16.VARICES: no, yes

17. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00

18. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250

19.SGOT: 13, 100, 200, 300, 400, 500,

20.ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0

21. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90

22. HISTOLOGY: no, yes

## 3.2.2 Data Preprocessing

After we've discovered, extracted, and purified the material we'll need for our use case, the next step is to understand it. These processes are used to preprocess the dataset. The first important problem is the enormous number of duplicated records in the Wisconsin (UCI Repository) dataset. We determined that around 68 percent and 65 percent of the entries in the Wisconsin (UCI Repository) dataset train and test sets are duplicated. Because the train set contains many redundant records, learning algorithms will be biased toward more frequent records, preventing them from being learnt. As a result, the purpose of this research is to offer a reliable diagnosis of hepatitis A and B, whether present or absent.

Because of its benefits in recognizing crucial characteristics from complicated hepatitis datasets, machine learning (ML) is generally recognized as the technique of choice in hepatitis pattern categorization and forecast modeling. Many research projects employ random subsets of the train set as test sets, which are how we got these statistics on both the train and test sets. As a result, they get a classification rate of roughly 98 percent utilizing simple machine learning algorithms.

Even utilizing the test set will yield in a minimum classification rate of 86 percent, making it difficult to compare Hepatitis disease detection algorithms because they all fall somewhere between 86 and 100 percent. We proposed a solution to the two problems mentioned in this study, resulting in new train and test sets comprised of chosen records from the full data set. In the data set supplied, none of the concerns occur. Furthermore, the train and test sets have enough records. This advantage allows the tests to be run on the complete collection rather than just a small sample at random. We used active learning to help us preprocess the dataset so that the deep learning system could grasp it. This is crucial since an accurate detection result requires a clean and non-redundant dataset in both the train and test sets.

The objective of this part is to ensure that the dataset is in a format that the machine learning algorithm can understand. We also want to avoid overfitting or under fitting the model with data, and we want to reduce the dimensionality of the dataset so that each stated parameter contributes at least equally to the proposed model's prediction.

# 3.3 Data Mining

Data mining is a knowledge discovery in data that helps in discovering hidden valuable knowledge, finding patterns, correlations within large datasets, and relationships within data. Alessandro (2019) posits that click baits have patterns and features and understanding these patterns and their relationship will help in detecting them thus preventing their effects. Click baits have certain features which make them bring traffic to their sites, these features have been stated above, and finding a paten with which they operate will help immensely with the use of data mining.

Data mining helps to discover data patterns automatically, predicts the likely outcomes, and create actionable information from large datasets and database. Data mining techniques include association, classification, clustering, prediction, sequential pattern, and decision trees. (Jared 2020) asserts that the primary aim of using data mining is to find anomalies, patterns, and correlations within large data sets to predict outcomes. The process of digging through data to discover hidden connections and predict future trends cannot be ignored in this century (Jared, 2020).

Data mining enables computers or the machines to make data-driven decisions rather than being explicitly programmed for carrying out a certain task. These programs or algorithms are designed in a way that they learn and improve over time when are exposed to new data.

Data mining techniques such as Machine learning algorithms are trained using a training data set to create a model. When new input data is introduced to the machine learning algorithm, it makes a prediction on the basis of the model.

The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine learning algorithm is deployed. If the accuracy is not acceptable, the Machine learning algorithm is trained again and again with an augmented training data set.

Alessandro (2019) conducted a survey on data mining techniques for detecting click baits, the result showed that an F1 of 0.93 was achieved using a data learning technique. Many data mining techniques have been exploited in clickbait detection like the Deep learning model and Support Vector Machine (SVM) and these models have improved click baits detection but there is still more room for research as patterns and more features keep arising (Alessandro,2019).

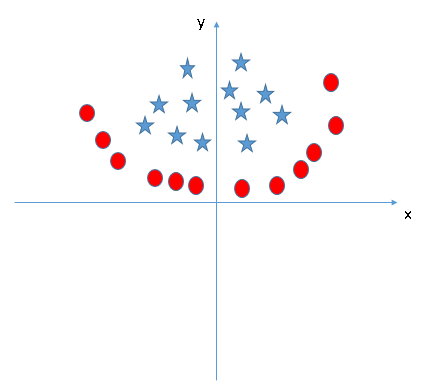
# 3.4 Model Design, Training, and Validation

In this approach, the creation of the model follows the selection and cleaning of the dataset. The machine learning methods Support Vector Machine, Random Forest Classifier, and Logistics Regression were used. They are algorithms for supervised data processing. The model is explained further in the section that follows.

## 3.4.1 Support Vector Machine

Support Vector machineisa form of supervised machine learning model or algorithm that learns appropriate data coding strategies, a process of segregating the two classes with a hyper-plane. It works well with a clear margin of separation and is effective in high-dimensional spaces.

It is effective in cases where the number of dimensions is greater than the number of samples and useful subset of training points in the decision function (called support vectors), so it is also memory efficient.



**Figure 3.1: Support Vector Machine**

The best hyperplane is that plane that has the maximum distance from both the classes, and this is the main aim of SVM. This is done by finding different hyperplanes which classify the labels in the best way then it will choose the one which is farthest from the data points or the one which has a maximum margin.

## 3.4.2 Random Forest Classifier

Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees. The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction.



**Figure 3.2: How random forest classifier works**

## 3.4.3 Logistics Regression

Logistic Regression is a classification technique used in machine learning. It uses a logistic function to model the dependent variable. The dependent variable is dichotomous in nature, i.e., there could only be two possible classes as a result, this technique is used while dealing with binary data. In logistic regression to map the predicted values to probabilities, sigmoid function is used. This function maps any real value into another value between 0 to 1. This function has a non-negative derivative at each point and exactly one inflection point.

# 3.5 *Evaluation of Classification Learning Methods*

The evaluation's goal is to analyze the model's performance on the input dataset. Hepatitis prediction is a classification problem, which means we have a set of characteristics or parameters called X and a label called Y that is a direct outcome of these attributes. As a result, our model must predict the label Y based on the input parameters X as shown in Table 3.1.

**Table 3.1: parameters and label table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute-1 | Attribute-2 | Attribute- 3 | … | Attribute-N | Label |
| X1 | X2 | X3 | … | XN | Y1 |
| X1 | X2 | X3 | … | XN | Y2 |
| X1 | X2 | X3 | … | XN | Y3 |
| X1 | X2 | X3 | … | XN | Y4 |
|  |  |  |  |  |  |
| XN | XN | XN | … | XN | YN |

To assess how well our model worked, we must look at how well it predicts the proper label given a set of input parameters. The performance measures employed here include the Accuracy, Sensitivity, Specificity, Precision, LogLoss and the Area under Receiver Operating Characteristics curve (AUC).

# 3.5.1 Confusion matrix for the predictions

The confusion matrix shows the ways in which classifier is confused when making a prediction and the types of errors that are being made (Brownlee, 2018c). A typical example of a confusion matrix for a binary class model is shown in Figure 3.3.

Table

Description automatically generated

**Figure 3.3 A confusion matrix**

**3.5.2 Basic terms related to confusion matrix.**

The following are the basics terms in a confusion matrix

**True Positives (TP)** - These are actually expected positive characteristics, showing that the value of the real class is 'yes,' very much like the value of the expected class.

**True Negatives (TN)-** These are unequivocally anticipated negative characteristics, inferring that the value of the real class is 'no' and the worth of the predicted class is also 'no.'

Values that occur when the actual and predicted classes do not match are known as False positives and negatives.

**False Positives (FP)-** This happens once the genuine class is "no" but the normal class is "yes."

**False Negatives (FN)-** This happens once the genuine class is "yes," yet the projected class is "no."

Our model's Accuracy, Precision, Recall, and F1 score can all be evaluated once these four parameters are known.

***Accuracy or Correct Classification Rate***: The accuracy of any given classifier is the ratio of response class labels that the classifier predicted correctly over the total number of predictions.

For a binary classifier, the classification rate is given by:

where TP = True Possitive; TN = True Nagative; FP = False Positive and FN = False Nagative.

In a class-imbalanced dataset, accuracy may not be an excellent method to evaluate classifiers; instead, we may use recall, precision, or the F1 score.

***True Positive Rate* (TPR)**: The True Positive Rate (TPR), often called the Sensitivity or Recall, is the ratio of the total number of samples that are correctly classified as having the response of interest (disease present) divided by the total number of samples that have the response of interest in the test data. Sensitivity is obtained as:

Sensitivity (TPR)

***True Negative Rate* (TNR)**: The True Negative (TNR), often called the Specificity, is the ratio of the total number of samples that are correctly classified as not having the response of interest (disease absent) divided by the total number of samples that do not have the response of interest in the test data. Specificity is obtained as:

Specificity (TNR)

***Precision***: The precision measures how often a classifier correctly predicts the response of interest (disease present). For instance, when the classifier predicts the hepatitis to be class A in a set of samples, how often is it correct? Precision can be expressed as follows:

***Log-Loss***: The Logarithmic Loss simply shortened as the Log-Loss is another metric to assess the goodness of a classifier. To calculate the Log-Loss, the classifier must predict each class's probability rather than merely predicting the class labels. The Log-Loss is calculated for each predicted probability and the average loss is then reported as the Log-Loss of the classifier. Mathematically, Log-Loss is expressed as:

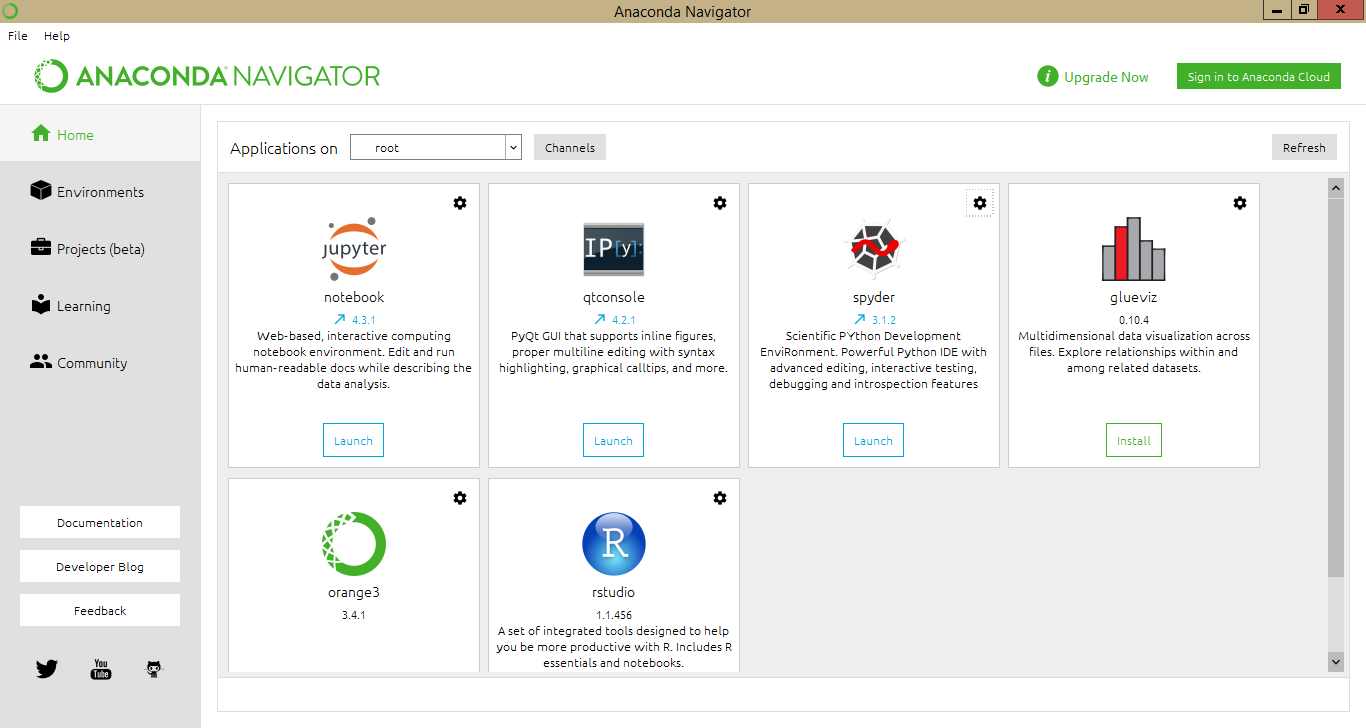
Log-Loss

where: is the sample size; is the true class label which is a binary indicator ( or ) that assumes indicator 0 if sample does not have the response of interest (disease absent) and assumes indicator 1 if sample has the response of interest (disease present); is the model’s predicted probability that sample is of class in .

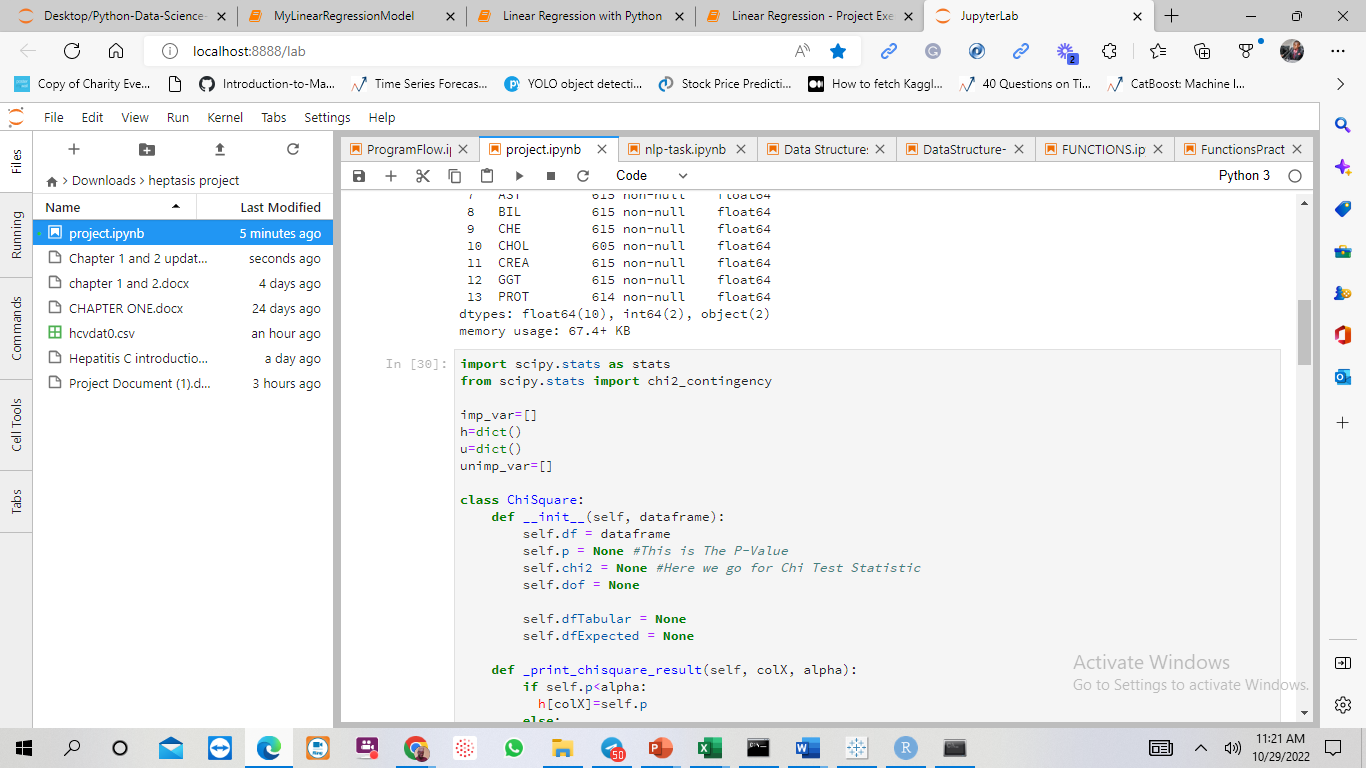
***Area under a ROC curve* (AUC)**:The Area Under the ROC Curve (AUC) metric is used to calculate the Area under the Reciever Operating Characteristics (ROC) curve. The AUC score is always bounded between zero and one, and a very poor classifier has an AUC of around 0.5 [37]. The AUC of a classifier represents the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation. Thus it is a useful metric even for datasets with highly unbalanced classes [37, 38]. Comparing the performance of different classifiers with the ROC curve is not easy [39. 40]. This is because no scalar value represents the expected performance [41].

# 3.6 Software Requirements

We used Anaconda as the development environment alongside Spyder which is a lunch tool for scientific analysis using Python in the Anaconda environment. The interface of Anaconda and Jupyter is shown in figure 3.5 – 3.6 below:



**Figure 3.4: The anaconda interface**



**Figure 3.5: The Jupyter notebook interface**

# 3.7 Hardware Requirement

The following computer configurations shall be used for this project.

* Pentium IV processor (minimum)
* 256MB-4GB RAM/ memory space
* 10GB hard disk space (minimum)
* CD ROM driver or DVD driver

# CHAPTER FOUR

# Implementation And Evaluation

* 1. **Experiment Analysis**

This portion of this study article dives into the details of the tests that were carried out. We present the results of the analyses and emphasize the work's prediction technique.

# Data and script availability

This project employed hepatitis datasets open to the public at the UCI (University of California, Irvine C.A) Machine Learning Repository and is available at <https://www.kaggle.com/datasets/amritpal333/hepatitis-c-virus-blood-biomarkers>. We used Jupyter lab Version 3.4.4. Packages used included the Pandas for data analysis, Matplotlib and Seaborn for data visualization and Sklearn for classification training. The raw data and scripts employed in this project are all available on GitHub via <https://github.com/Sharpsharp007/Hepatitis-A-and-B-predictions>.

This dataset in Figure 4.1 will be used to train various models such as the Support Vector Machine, Random Forest Classifier, and Logistics Regression in this project.

Table

Description automatically generated with medium confidence

**Figure. 4.1: A glimpse of the hepatitis data**

# 4.3 Exploratory Data Analysis

The data have twelve (12) features, representing the bio-clinical variables obtained on hepatitis disease patients as shown in Figure 4.1. These features were used in the various models in this study to predict whether a patient has hepatitis A or B.

The hepatitis dataset has 615 patients, out of which 533 (86.7%) patients had hepatitis A, and 82 (13.3%) patients have hepatis B as shown in Figure 4.2

Chart, sunburst chart

Description automatically generated

**Figure 4.2: Distribution of hepatitis among patients**

We looked at the numeric values in the dataset to get a sense of what it was like. The output of the describe function on the dataset is shown in Figure 4.3.

Table

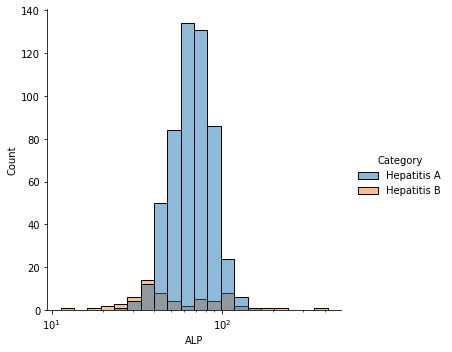
Description automatically generated

**Figure 4.3: Descriptive statistics**

We examined the age distribution of the dataset using a histogram as shown in Figure 4.4 and age is symmetric.



**Figure 4.4: Age distribution**

****

**Figure 4.5: Distribution of hepatitis categories over ALP**

**A picture containing text, scoreboard

Description automatically generated**

**Figure 4.6: Heat map of the dataset's correlations to highlight how they affect the prediction's outcome.**

Every cell in the correlation matrix typifies a 'correlation coefficient' between the two factors that symbolize the cell's row and column. A connection coefficient is referred to as a number that demonstrates how solid a relationship exists between two factors. There are various correlation coefficients to consider.

Pearson's coefficient ρ (condensed as (rho)) is the most broadly used. It's evaluated by obtaining the quotient of the covariance between two factors and the product of the two factors' standard deviations.

# 4.3.1 Formula for correlation coefficient between two variables

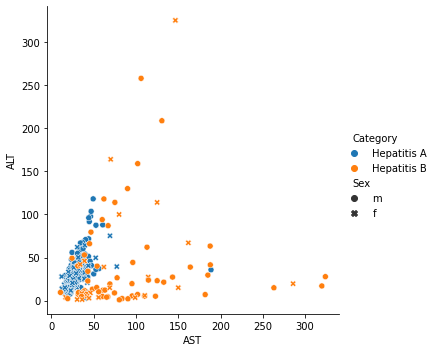
Where COV(X, Y) is described as the "expected value of the product of X and Y's deviations from their particular means."

* + - The value of ρ ranges from -1 to +1.
    - Values around +1 suggest that X and Y have a strong positive relationship, whereas values near -1 indicate that X and Y have a strong negative relationship.
    - Values close to 0 indicate that there is no link between X and Y.

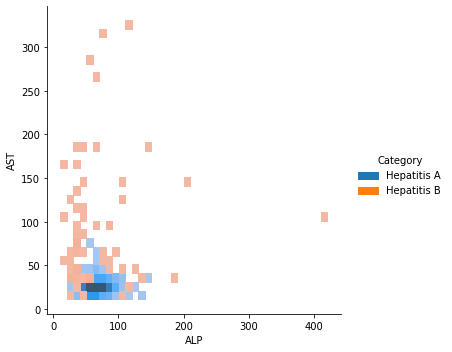
There appear to be some relationships in Biomarker levels, as seen by the heat map. In figure 4.4, we display the parameter regression plot along with the relation plot and the distribution plot.



**Figure 4.7: the parameter regression plot along with the relation plot and the distribution plot.**

****

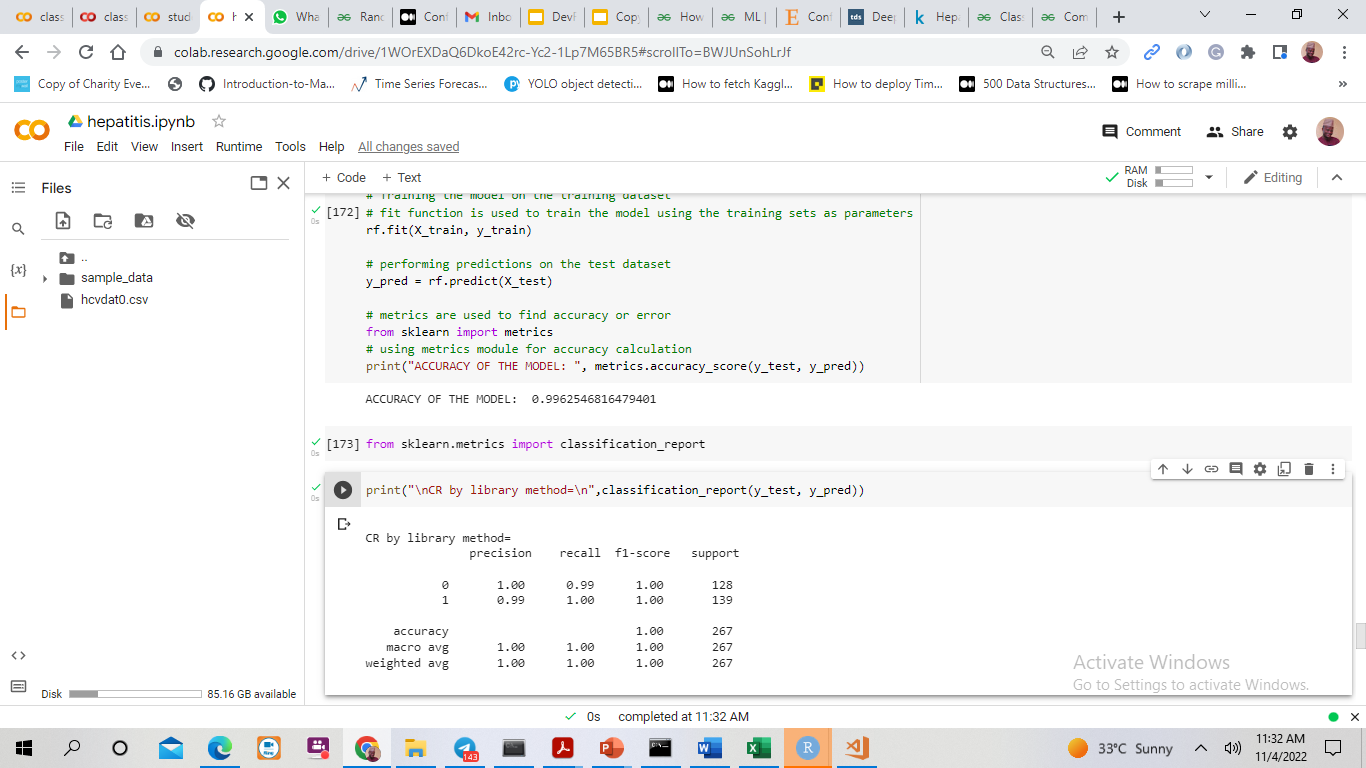
**Figure 4.8 The categories over sex with AST**

****

**Figure 4.9 The categories over ALP.**

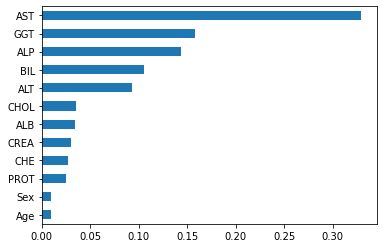
## 4.4 Training and testing of the models

In the implementation of training and testing of the model, 70% of the samples were utilized for training and 30% of the samples were used for testing. Steps were taken to ensure that the individuals (Patients) that were selected for training were not reused during testing to generalize the task of classification and perform satisfactorily when testing new patients.



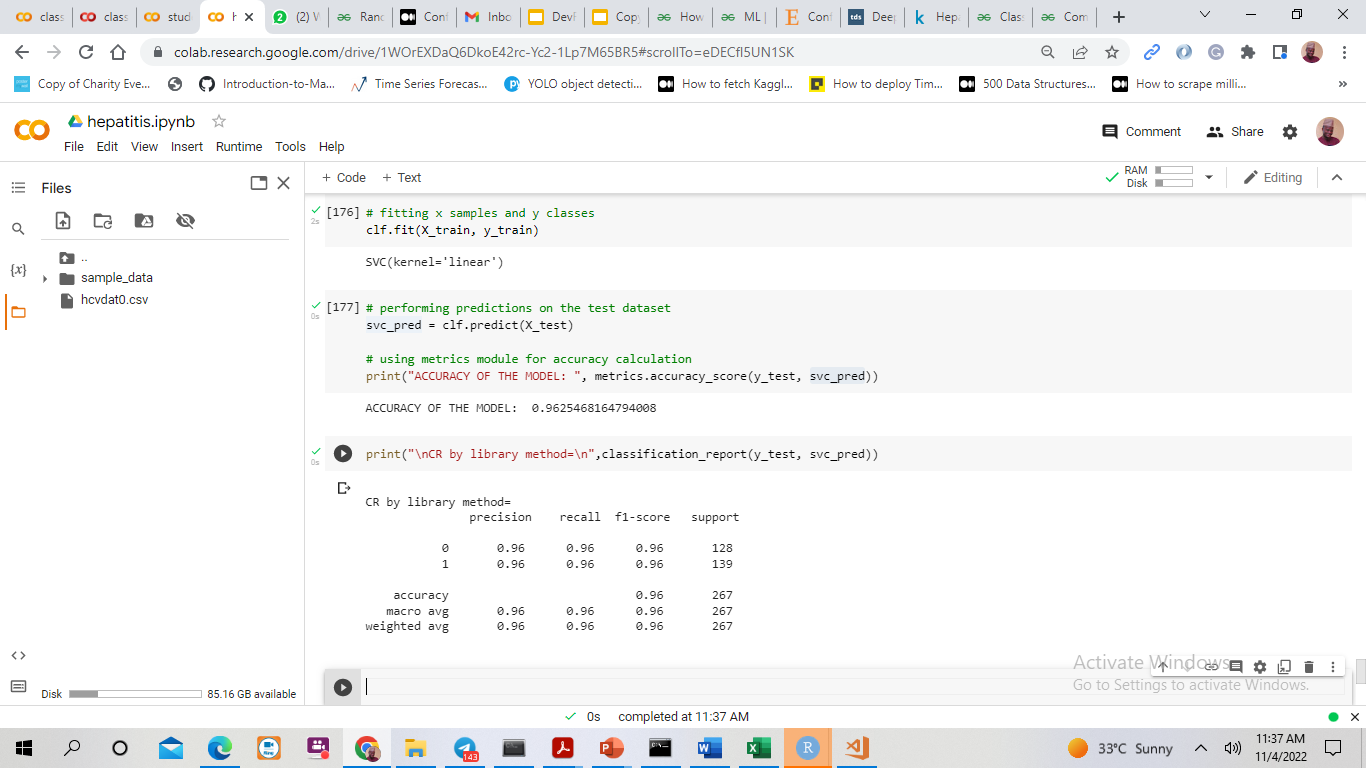
**Figure 4.10: Result of Precision, recall, f1 score for random forest classifier**

The best results for each model execution are recorded in the table above.



**Figure 4.11 Important features for the random forest classifier**

The result of the Support Vector Classifier



**Fig 4.12 The result of precision, recall, f1 score for the Support Vector classifier**

From the three compared models, the random forest classifier gave 100% performance on the dataset used, followed by Support Vector Classifier with 96% and logistics regression with 97%.

# References

1. Andino R, Domingo E. Viral quasispecies. Virology. 2015 May;479-480:46-51. doi: 10.1016/j.virol.2015.03.022. Epub 2015 Mar 29. PMID: 25824477; PMCID: PMC4826558.
2. Caligiuri, P., Cerruti, R., Icardi, G., &Bruzzone, B. (2016). Overview of hepatitis B virus mutations and their implications in the management of infection. *World journal of gastroenterology*, *22*(1), 145–154. <https://doi.org/10.3748/wjg.v22.i1.145>
3. CW Spearman, M Afihene, R Ally, B Apica, Y Awuku, L Cunha. Journal of viral hepatitis 22, 26-45, 2015.
4. Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627-9636).
5. Jeong J., Kang H.-M., Lee E.-K., Song B.-M., Kwon Y.-K., Kim H.-R., Choi K.-S., Lee Y.-J. Highly pathogenic avian influenza virus (H5N8) in domestic poultry and its relationship with migratory birds in South Korea during (2014) Veterinary *Microbiology*, 173 (3-4), pp. 249-257.
6. Yarasuri, V. K., Indukuri, G. K., & Nair, A. K. (2019). Prediction of Hepatitis Disease Using Machine Learning Technique. 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC), 265â€“269. <https://doi.org/10.1109/I-SMAC47947.2019.9032585>
7. X Tian, Y Chong, Y Huang, P Guo, M Li Using machine learning algorithms to predict hepatitis B surface antigen seroclearance 2019
8. [BA Wintle](https://scholar.google.com/citations?user=N6QmJeMAAAAJ&hl=en&oi=sra), [O Broennimann](https://scholar.google.com/citations?user=NGCG1bMAAAAJ&hl=en&oi=sra), [M Austin](https://scholar.google.com/citations?user=NqOdAxwAAAAJ&hl=en&oi=sra) predicting species distributions for conservation decisions 2013
9. Samuel, A. L. (1988). Some studies in machine learning use the game of checkers. II—recent progress. *Computer Games I*, 366-400.
10. Nguyen, Lam M., et al. "SARAH: A novel method for machine learning problems using a stochastic recursive gradient." *International Conference on Machine Learning*. PMLR, 2017.
11. Zinszner, H., Kuroda, M., Wang, X., Batchvarova, N., Lightfoot, R. T., Remotti, H., ... & Ron, D. (1998). CHOP is implicated in programmed cell death in response to impaired function of the endoplasmic reticulum. *Genes & development*, *12*(7), 982-995.
12. Boland, Angela, Gemma Cherry, and Rumona Dickson, eds. "Doing a systematic review: A student's guide." (2017).
13. Jayathilaka, Wanasinghe Arachchige Dumith Madushanka, et al. "Significance of nanomaterials in wearables: a review on wearable actuators and sensors." *Advanced Materials* 31.7 (2019): 1805921.
14. Aswathy, M. A., &Jagannath, M. (2017). Detection of breast cancer on digital histopathology images: Present status and future possibilities. *Informatics in Medicine Unlocked*, *8*, 74-79.Bandyopadhyay, Samir & Dutta, Shawni. (2020). A Voting Ensemble Approach for Hepatitis Disease Detection. International Journal For Research in Applied Sciences and Biotechnology. 7. 56-62. 10.31033/ijrasb.7.5.6.
15. Tillmann, A. M. 2015. "On the Computational Intractability of Exact and Approximate Dictionary Learning". IEEE Signal Processing Letters. 22 (1): 45–49. arXiv:1405.6664. Bibcode:2015ISPL...22...45T. doi:10.1109/LSP.2014.2345761.
16. Boström, Christoffer, et al. "Seascape ecology of coastal biogenic habitats: advances, gaps, and challenges." *Marine ecology progress series* 427 (2011): 191-217.
17. Tian, Zhi, et al. "Fcos: Fully convolutional one-stage object detection." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
18. Louis, David N., et al. "The 2016 World Health Organization classification of tumors of the central nervous system: a summary." *Acta neuropathologica* 131.6 (2016): 803-820.
19. Vos, T., Allen, C., Arora, M., Barber, R. M., Bhutta, Z. A., Brown, A., ... &Boufous, S. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The lancet*, *388*(10053), 1545-1602.
20. Mack CL, Adams D, Assis DN, Kerkar N, Manns MP, Mayo MJ, Vierling JM, Alsawas M, Murad MH, Czaja AJ. Diagnosis and management of autoimmune hepatitis in adults and children: 2019 practice guidance and guidelines from the American Association for the Study of Liver Diseases. Hepatology. 2020 Aug;72(2):671-722.
21. Huber, Lawrence, et al. "Orchestrating high-throughput genomic analysis with Bioconductor." *Nature methods* 12.2 (2015): 115-121.
22. Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Feature subset selection problem using wrapper approach in supervised learning." *International journal of Computer applications* 1.7 (2010): 13-17.
23. Kimball, Ralph, and Richard Merz. "The data webhouse toolkit: Building the Web‐enabled data warehouse." *Industrial Management & Data Systems* (2000).
24. Liu, Haining, and M. E. Zarki. "A synchronization control scheme for real-time streaming multimedia applications." *Packet Video*. Vol. 2003. 2003.
25. Troyanskaya, Olga, et al. "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17.6 (2001): 520-525.
26. Andronie, M. (2010). Economic process optimization deployed by companies with complex activity production–trade–services using informatics systems with integrated data mining techniques. *Annals of SpiruHaret University*, *1*(1), 77-84.
27. Koza, John R., et al. "Automated design of both the topology and sizing of analog electrical circuits using genetic programming." *Artificial intelligence in design’96*. Springer, Dordrecht, 1996. 151-170.
28. Mitchell, T. 1997. Machine Learning. McGraw Hill. p. 2. ISBN 978-0-07-042807-2.
29. Harnad, Steva. 2008. "The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence", in Epstein, Robert; Peters, Grace, The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer, Kluwer
30. Bishop, C. M. 2006., Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2 breast cancer histopathological image classiﬁcation,” IEEE Transactions
31. Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, *66*(4), 486-501.
32. de Paiva Penha, Deyvison, and Adriana Rosa Garcez Castro. "Home appliance identification for NILM systems based on deep neural networks." *Int. J. Artif. Intell. Appl* 9 (2018): 69-80.
33. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20.3 (1995): 273-297.
34. Lambers, A., Rieger, B., Kop, A., D’Alessandro, P., & Yates, P. (2019). Implant fracture analysis of the TFNA proximal femoral nail. *JBJS*, *101*(9), 804-811.
35. Zhang, Lin, et al. "FSIM: A feature similarity index for image quality assessment." *IEEE transactions on Image Processing* 20.8 (2011): 2378-2386.
36. Oishi et al. advocated using an Effective Deep Neural Network (2020).
37. Yarasuri, Indukuri, and Nair's paper "Prediction of Hepatitis Disease Using Machine Learning Technique" (2019).
38. PinarYildirim “Filter-Based Feature Selection Methods for Risk Prediction in Hepatitis Disease” (2015).
39. Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." *IEEE/ACM Transactions on audio, speech, and language processing* 22.10 (2014): 1533-1545.