# APPLICATION OF MACHINE LEARNING ALGORITHMS TO A WELL DEFINED CLINICAL PROBLEM: LIVER DISEASE

*Dissertation submitted in fulfilment of the requirements for the Degree of*

**MASTER OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**


By

**SAKSHI**

**11502963**

Supervisor

**ASST. PROF AMAN SINGH**




**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)


April, 2017

April, 2017

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

**TOPIC APPROVAL PERFORMA**

School of Computer Science and Engineering

**Program :** P172::M.Tech. (Computer Science and Engineering) [Full Time]

**COURSE CODE :** CSE546 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0018

**Supervisor Name :** Aman Singh **UID :** 16826 **Designation :** Assistant Professor

**Qualification :** M.Tech **Research Experience :** 5 yrs

| SR.NO. | NAME OF STUDENT | REGISTRATION NO | BATCH | SECTION | CONTACT NUMBER |
|---|---|---|---|---|---|
| 1 | Sakshi | 11502963 | 2015 | K1518 | 9988851303 |

**SPECIALIZATION AREA :** Networking and Security **Supervisor Signature:** Aman 16826 21/04/17

**PROPOSED TOPIC :** Application of machine learning algorithms in liver disease diagnosis

| Qualitative Assessment of Proposed Topic by PAC | | |
|---|---|---|
| **Sr.No.** | **Parameter** | **Rating (out of 10)** |
| 1 | Project Novelty: Potential of the project to create new knowledge | 7.50 |
| 2 | Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students. | 6.75 |
| 3 | Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program. | 7.00 |
| 4 | Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills. | 8.00 |
| 5 | Social Applicability: Project work intends to solve a practical problem. | 7.50 |
| 6 | Future Scope: Project has potential to become basis of future research work, publication or patent. | 7.25 |

| PAC Committee Members | | |
|---|---|---|
| PAC Member 1 Name: Prateek Agrawal | UID: 13714 | Recommended (Y/N): NA |
| PAC Member 2 Name: Pushpendra Kumar Pateriya | UID: 14623 | Recommended (Y/N): Yes |
| PAC Member 3 Name: Deepak Prashar | UID: 13897 | Recommended (Y/N): Yes |
| PAC Member 4 Name: Kewal Krishan | UID: 11179 | Recommended (Y/N): Yes |
| PAC Member 5 Name: Anupinder Singh | UID: 19385 | Recommended (Y/N): NA |
| DAA Nominee Name: Kanwar Preet Singh | UID: 15367 | Recommended (Y/N): Yes |

**Final Topic Approved by PAC:** **Application of machine learning algorithms in liver disease diagnosis**

**Overall Remarks:** Approved

**PAC CHAIRPERSON Name:** 11024::Amandeep Nagpal **Approval Date:** 05 Mar 2017

4/20/2017 12:58:13 PM

# ABSTRACT

Correct diagnosis of a disease is a very crucial task in medical domain. Liver illness is the most hazardous ailment that influences a large number of individuals consistently and ends man's life. Machine learning algorithms have been extensively used by doctors for diagnosing liver disease so as to enhance efficiency of medical diagnosis. This study accordingly employs various machine learning algorithms on distinct liver disease datasets to evaluate the diagnostic performances in terms of different parameters, to integrate dimensionality reduction and optimization techniques with the selected classification algorithms for analyzing the variation in results, to find the best classification models for liver diseases, to analyze the merits and demerits of applied techniques. This study includes algorithms like linear and diagonal linear discriminant analysis (LDA and DLDA), quadratic and diagonal quadratic discriminant analysis (QDA and DQDA), support vector machine (SVM), K- nearest neighbor (KNN) and case-based reasoning algorithm (CBR). To analyze variation in results another dimensionality reduction and optimization techniques are used to develop integrated models. By analyzing the results of these algorithms on different datasets, it is concluded that one particular algorithm can't show the best results for all types of datasets as no algorithm is perfect. The performance of an algorithm totally depends on the dataset type and structure, its number of observations, its dimensions and the decision boundary.

***Keywords:*** *linear and diagonal linear discriminant analysis, quadratic and diagonal quadratic discriminant analysis, support vector machine, K- nearest neighbor, case-based reasoning algorithm, genetic algorithm, principal component analysis.*

# DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "**APPLICATION OF MACHINE LEARNING ALGORITHMS TO A WELL DEFINED CLINICAL PROBLEM: LIVER DISEASE**" in partial fulfilment of the requirement for the award of Degree for Master of Technology in **Computer Science and Engineering** at **Lovely Professional University, Phagwara, Punjab** is an authentic work carried out under supervision of my research supervisor **Mr. Aman Singh**. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

**Date:**

**Name:** Sakshi

**Regn No:** 11502963

# SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**APPLICATION OF MACHINE LEARNING ALGORITHMS TO A WELL DEFINED CLINICAL PROBLEM: LIVER DISEASE**", submitted by **Sakshi** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

_____

Signature of Supervisor

**Name:** Mr. Aman Singh

**Date:**

**Counter Signed by:**

1) **Concerned HOD:**
   HoD's Signature: _____
   HoD Name: _____
   Date: _____

2) **Neutral Examiners:**

   **External Examiner**

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   **Internal Examiner**

   Signature: _____

   Name: _____

   Date: _____

# ACKNOWLEDGEMENT

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

I would like to convey my most heartfelt and sincere gratitude to my mentor **Mr. Aman Singh** for his valuable guidance, advice, understanding and supervision throughout the development of this dissertation study. His willingness to motivate me contributed tremendously to achieve the goal successfully. I would like to thank to the **Project Approval Committee members** for their valuable comments and discussions.

I owe my thanks to my family and friends for their consistent moral support. They are ones who have encouraged me at every step of my life.

**Name:** Sakshi
**Regn No:** 11502963

# TABLE OF CONTENTS

ABSTRACT ..... iii
DECLARATION STATEMENT ..... iv
SUPERVISOR'S CERTIFICATE ..... v
ACKNOWLEDGEMENT ..... vi
LIST OF TABLES ..... ix
LIST OF FIGURES ..... x
1 INTRODUCTION ..... 1
1.1 LIVER-AN INTRODUCTION ..... 1
1.1.1 Anatomy of Liver ..... 1
1.1.2 Functions of Liver ..... 2
1.2 OVERVIEW OF LIVER DISEASE ..... 3
1.2.1 Types of Liver Disease ..... 4
1.2.2 Causes of Liver Disease ..... 5
1.2.3 Risk Factors of Liver Disease ..... 6
1.2.4 Symptoms of Liver Disease ..... 6
1.2.5 Diagnosis of Liver Disease ..... 6
1.3 INTRODUCTION TO MACHINE LEARNING APPROACHES FOR LIVER DISEASE CLASSIFICATION ..... 6
1.3.1 Machine Learning ..... 7
1.3.2 Classification ..... 7
1.3.3 Single and Hybrid Machine Learning Approach ..... 8
1.4 THESIS ORGANIZATION ..... 9
2 REVIEW OF LITERATURE ..... 10
3 RATIONALE AND SCOPE OF THE STUDY ..... 19
3.1 RATIONALE OF THE STUDY ..... 19
3.2 SCOPE OF THE STUDY ..... 19
4 PRESENT WORK ..... 21
4.1 PROBLEM FORMULATION ..... 21
4.2 OBJECTIVES OF STUDY ..... 22
4.3 RESEARCH METHODOLOGY ..... 23
4.3.1 Proposed Methodology I ..... 23

4.3.2 Proposed Methodology II ..... 31
5 RESULTS AND DISCUSSION ..... 39
5.1 EXPERIMENTAL WORK ..... 39
5.1.1 Tools description/ Simulation work ..... 39
5.1.2 Techniques and Methodologies used ..... 40
5.2 DATA ANALYSIS AND INTERPRETATION ..... 41
5.3 PERFORMANCE EVALUATION ..... 42
5.3.1 Performance Evaluation Parameters ..... 42
5.3.2 Experimental Results of Methodology 1 ..... 44
5.3.3 Experimental Results of Methodology II ..... 49
6 CONCLUSION AND FUTURE SCOPE ..... 54
7 REFERENCES ..... 56
8 APPENDIX ..... 60
9 PUBLICATONS ..... 61

# LIST OF TABLES

**Table 5.1** Feature details of BUPA liver disorder dataset. ..... 41
**Table 5.2** Feature details of Liver damage dataset. ..... 42
**Table 5.3** Feature details of Hepatitis dataset. ..... 42
**Table 5.4** The comparison results of single models for BUPA liver disorder dataset. 45
**Table 5.5** The comparison results of single models for Liver damage dataset. ..... 45
**Table 5.6** The comparison results of single models for Hepatitis dataset. ..... 45
**Table 5.7** The comparison results of PCA-based integrated models for BUPA liver disorder dataset. ..... 46
**Table 5.8** The comparison results of PCA-based integrated models for Liver damage dataset. ..... 47
**Table 5.9** The comparison results of PCA-based integrated models for Hepatitis dataset. ..... 47
**Table 5.10** Case details of different datasets. ..... 51
**Table 5.11** CBR Accuracy details of different datasets. ..... 51
**Table 5.12** GA-CBR Accuracy details of different datasets. ..... 51

# LIST OF FIGURES

**Figure 4.1** The Overall System design for the diagnosis of liver diseases ..... 23
**Figure 4.2** A CBR-GA based integrated model for diagnosis of liver diseases..... 31

# CHAPTER 1
# INTRODUCTION

## 1.1 LIVER-AN INTRODUCTION

The liver is the biggest inner organ. It is reddish chestnut, weighs roughly three pounds (in the grown-up male) and is about the extent of a football. It is situated behind the ribcage on the upper right half of the guts. The liver has the remarkable capacity to recover its own tissue—as much as seventy five percent of the liver can be lost, and the organ can become back or grow to its unique size inside a few weeks. This permits individuals who need transplants to get part of a living or perished contributor's liver.

The liver is isolated into four projections; these are thus made out of numerous lobules which contain the hepatocytes, or working liver cells. The liver has a broad blood supply—around 1½ quarts of blood course through it consistently. It gets oxygen-rich blood from the hepatic vein. The entrance vein conveys blood containing supplements, poisons, and different substances retained from the digestion tracts to the liver.

### 1.1.1 Anatomy of Liver

The liver is a generally triangular organ that reaches out over the whole stomach hole just inferior to the diaphragm. A large portion of the liver's mass is situated on the right half of the body where it slides poorly toward the right kidney. The liver is made of delicate, pinkish-cocoa tissues typified by a connective tissue case. This case is further secured and fortified by the peritoneum of the stomach cavity, which ensures the liver and holds it set up inside the mid-region.

There are two particular sources that supply blood to the liver: Oxygenated blood streams in from the hepatic supply route and supplement rich blood streams in from the hepatic entry vein. The liver comprises of two primary lobes, both of which are comprised of 8 fragments. The portions are comprised of a thousand lobules. The lobules are associated with little pipes that interface with bigger pipes to at last shape the normal hepatic pipe. The regular hepatic pipe transports bile created by the liver

cells to the gallbladder and duodenum (the initial segment of the small digestive tract).

### 1.1.2 Functions of Liver

The liver is in charge of somewhere in the range of 500 real capacities. It assumes a part in assimilation, sugar and fat digestion system, and the body's resistant safeguard. It forms practically everything a man eats, inhales, or ingests through the skin. Around 90% of the body's supplements go through the liver from the digestion systems. Liver converts sustenance into vitality, stores supplements, and delivers blood proteins. The liver additionally goes about as a channel to expel destructive substances from the blood. In the creating baby, platelets are delivered in the liver.

***Digestion***

The liver assumes a critical part in the assimilation and handling of nourishment. Liver cells produce bile, a greenish-yellow liquid that guides the processing of fats and the ingestion of fat-dissolvable supplements. Bile is conveyed to the small digestive tract through the bile conduit; when there is no sustenance to process, additional bile is put away in a little organ called the gallbladder situated underneath the liver. By-items from the breakdown of medications and poisonous substances prepared by the liver are conveyed in the bile and discharged from the body [1]. A man with a harmed liver may encounter debilitated bile generation and stream. When this happens, the body will be unable to appropriately retain supplements. Liver cells likewise change over heme (a segment of hemoglobin that is discharged when red platelets are separated) into bilirubin. At the point when the liver is harmed, bilirubin may develop in the blood, creating jaundice (yellowing of the skin and whites of the eyes).

***Metabolism***

The liver does numerous metabolic capacities, giving the body the vitality it needs. It directs the creation, stockpiling, and arrival of sugar, fats, and cholesterol. At the point when nourishment is eaten, the liver believers (glucose) into glycogen, which is put away for later utilize. At the point when vitality is required, the liver proselytes glycogen once more into glucose in a procedure called gluconeogenesis. The liver manages the capacity of fats by changing over amino acids from processed nourishment into unsaturated fats, for example, triglycerides; when the body does not have enough sugar, the liver converts unsaturated fats into ketones, which can be

utilized for fuel. The liver additionally controls the creation, digestion system, and discharge of cholesterol, which is a vital segment of cell layers and certain hormones.

***Storage***

The liver stores a few supplements, including vitamins A, D, B-9 (folate), and B12. It likewise stores press and assumes a part in changing over iron into heme, a segment of hemoglobin (the oxygen-conveying particle in red platelets) [1].

## 1.2 OVERVIEW OF LIVER DISEASE

Liver diseases represent a major health burden worldwide. It is rising day by day and is not discovered easily in its initial stage as liver can work properly even when it is partially damaged. Liver illness is affecting the large group of people. It kills a greater number of individuals than diabetes and street passings joined. They don't consider a man's own particular hazard elements, for example, family history, practices, or disease screenings. According to the most recent WHO information distributed in May 2014 liver disease deaths in India achieved 216,865 or 2.44% of aggregate passings. The age balanced death rate is 21.96 for each 100,000 of population ranks India 61 in the world. In Mumbai alone, about 2,000 patients die yearly from liver disappointment or liver tumor, and 300 are sitting tight for a liver transplant at any time.

Liver sickness – including cirrhosis, liver cancer and hepatitis is the twelfth driving reason for death in the United States. Consistently roughly 15,000 Americans pass on from liver tumor or endless liver ailment related with viral hepatitis. Some statistics about liver cancers are there which estimates that around 40,710 new instances of essential liver cancer will be analyzed in the United States in 2017 and around 28,920 individuals will die because of liver cancer in the United States in 2017. As per national measurements in the UK, liver illnesses have been positioned as the fifth most regular reason for death [2]. There are an expected 30,000 individuals living with cirrhosis in the UK and no less than 7,000 new cases being analyzed every year. Most of the symptoms of liver disease are similar to other fever related ailments which result in the wrong diagnosis of disease. As other diseases dominate the liver disease due to which it is unable to identify. It is a very tough job for the physicians to recognize the disease from common symptoms.

### 1.2.1 Types of Liver Disease

There are various types of liver disease which is described as follows:

*a. Alcoholic liver disease*

Alcoholic liver disease is liver harm caused by high alcohol consumption. Liver disease may start to create after an "edge" measurement of liquor has been expended. Everyone who consumes very large quantity of alcohol will have some symptom of liver injury. The suspicion that alcoholic liver disease dependably advances directly from fatty liver, to alcoholic hepatitis and at last to cirrhosis is not right.

*b. Liver Cirrhosis*

Cirrhosis is a liver condition in which there is damage to liver that is not reversible. The fundamental causes are extreme liquor utilization, viral hepatitis B and C, and fatty liver sickness - be that as it may, there are numerous conceivable causes. Individuals with cirrhosis may create jaundice, itching and outrageous tiredness [3]. For cirrhosis to grow long term, persistent harm to the liver needs to happen.

*c. Non-Alcoholic Fatty Liver Disease*

Non-alcoholic fatty liver disease (NAFLD) is the development of additional fat in liver cells that is not brought about by liquor. It is typical for the liver to contain some fat. Nonetheless, if over 5% - 10% percent of the liver's weight is fat, then it is known as a fatty liver (steatosis).

*d. Liver Cancer (Hepatocellular Carcinoma)*

When healthy cells change and grow out of control then cancer starts and, framing a mass called a tumor. A tumor can be destructive or amiable. A dangerous tumor is harmful, which means it can develop and spread to different parts of the body. A considerate tumor implies the tumor can develop however won't spread. It is more typical for growth that began in another piece of the body to spread to the liver. This is not liver tumor, but instead metastatic disease of another organ.

*e. Hepatitis*

Hepatitis is an injury to liver due to inflammation of liver cells. The condition can act naturally constraining or it can advance to liver cancer, cirrhosis or fibrosis. Recognized reason for hepatitis in the world is Hepatitis Virus yet different

contaminations, harmful substances (e.g. liquor, certain medications), and immune system sicknesses can likewise bring about hepatitis [4].

### 1.2.2 Causes of Liver Disease

Liver sickness happens when substantial parts of liver get to be distinctly harmed and is no more extended ready to work. Liver ailment has many causes.

*a. Contamination*

Parasites and diseases can spoil the liver, bringing on exacerbation and that reduces liver limit. The diseases that cause liver damage can be spread through blood or semen, degraded food or water, or close contact with a man who is sullied. The most generally perceived sorts of liver ailment are hepatitis diseases, including: Hepatitis A, Hepatitis B and Hepatitis C.

*b. Immune system abnormality*

Sicknesses in which your insusceptible structure strikes certain parts of your body (safe framework) can impact your liver. Instances of insusceptible framework liver ailments include: Primary sclerosing cholangitis, essential biliary cirrhosis and immune system hepatitis.

*c. Hereditary qualities*

An unusual quality acquired from either of your folks can bring about different substances to develop in your liver, bringing about liver harm. Hereditary liver maladies include: Wilson's ailment, Hyperoxaluria and oxalosis and Hemochromatosis.

*d. Malignancy and different developments*

Cases include: Liver adenoma, Liver malignancy and Bile channel disease and so on.

*e. Other*

Extra, regular reasons for liver sickness include: Chronic liquor mishandle and fat collecting in the liver (nonalcoholic greasy liver illness).

### 1.2.3 Risk Factors of Liver Disease

Components that may build your danger of liver ailment include:

Overwhelming liquor utilize, tattoos or body piercings, unprotected sex, infusing drugs utilizing shared needles, injecting drugs using shared needles, introduction to specific chemicals or poisons, diabetes, obesity, large amounts of triglycerides in your blood.

### 1.2.4 Symptoms of Liver Disease

The underlying side effects of liver disappointment are frequently ones that can be because of any number or conditions. Along these lines, liver disappointment might be at first hard to analyze. Early indications include: nausea, diarrhea, fatigue, swollen abdomen, sleepiness, coma, jaundice, bleeding easily, mental disorientation, loss of appetite, vomiting, dark urine color, tendency to bruise easily, itchy skin, irritated skin, fluid retention, exhaustion, stomach pain, muscle or joint pain, skin rash, hair loss, depression and unusual agitation etc [5].

### 1.2.5 Diagnosis of Liver Disease

Finding the cause and degree of liver harm is imperative in controlling treatment. Your specialist is probably going to begin with a wellbeing history and exhaustive physical examination. Your specialist may then prescribe:

*a.* *Imaging tests:* MRI, ultrasound and CT scan can demonstrate liver harm.
*b.* *Blood tests:* A gathering of blood tests called liver capacity tests can be utilized to analyze liver ailment. Other blood tests should be possible to search for particular liver issues or hereditary conditions.
*c.* *Tissue analysis:* Expelling a tissue test (biopsy) from your liver may help analyze liver infection. Liver biopsy is frequently done utilizing a long needle embedded through the skin to remove a tissue test. It is then broke down in a lab.

## 1.3 INTRODUCTION TO MACHINE LEARNING APPROACHES FOR LIVER DISEASE CLASSIFICATION

Liver disease is rising day by day and is not discovered easily in its initial stage as liver can work properly even when it is partially damaged. It is important to diagnose

liver disease early which can increase the patient's survival rate. Expert physicians are required for various examination tests to diagnose the liver disease, but it cannot assure the correct diagnosis. It is very tough job for the expert physicians to recognize the disease from common symptoms. Most of the symptoms are similar to other fever related ailments which result the wrong diagnosis of disease. As the other diseases dominate the liver disease due to which it is unable to identify. Computer-aided diagnosis is needed for correct prediction of liver disease and it also helps to deal with tremendous and cumbersome data. Research interest is growing in the field of machine learning and knowledge discovery in order to traverse knowledge in detailed volume. Data stored in databases contains valuable hidden knowledge which helps to enhance decision making.

### 1.3.1 Machine Learning

Machine learning is an artificial intelligence branch in which a computer program learns from the experience and the performance if its performance measure improves with the experience. E.g. we can train a machine learning system to diagnose whether the person suffers from liver disease or not.

Why machine learning is important-

a. Some tasks can be defined well only by examples.
b. Machine learning helps us to find the hidden correlations and relationships from the large amount of data.
c. Gives better results for prediction and model generation.
d. Environments may change sometimes.
e. Some problems with huge amount of knowledge too hard for humans to be described.

### 1.3.2 Classification

Classification is a process which consists of predicting a specific result in presence of given input information. There is a requirement of one training set which consists of a set of attributes and the respective outcomes called the goal attribute to determine the specified output. The algorithm tries to find connections between the attributes that would make it conceivable to anticipate the result. Then an unknown dataset is provided to the algorithm which consists of same set of attributes except for the

attribute which is not known. The algorithm analyzes the input information and the produces the respective outcome. In this there is a database of medical information related to liver disease where the prediction attribute is whether the patient suffers from liver disease or not.

Classification is of two types: Supervised Classification and Unsupervised Classification. Supervised Classification is one of the main methods to extract knowledge from databases where set of training examples are known previously and in unsupervised classification training examples are not known previously. Actually, Classification is a dual process which consists two phases. One is Training phase where with the help of classifier algorithm, training dataset trains the classifier. The other is Testing phase where testing of classifier is done to analyze its performance using different samples of the test set. Prediction accuracy is a criterion to evaluate the performance of classifier. Classification accuracy describes the percentage of instances which are correctly classified.

### 1.3.3 Single and Hybrid Machine Learning Approach

Single and hybrid machine learning approaches are used to predict whether the patient suffers from liver disease or not. In single machine learning approach, an algorithm is applied to the liver disease dataset and accuracy is calculated for performance evaluation. In hybrid machine learning approach, a single approach is integrated with other algorithms or some dimensionality reduction techniques in order to analyze the variation in results. Various classification algorithms are there which include SVM, discriminant analysis and nearest neighbor algorithms, decision tree etc. These classification algorithms are applied on different small or large medical datasets. The task of learning from scanty datasets is an arduous task. Some datasets contain too many attributes but to select an adequate subset of attributes or features is a significant question. To select an effective subset of attributes, two dimension reduction techniques are there – one technique is to reduce the dimensions by selecting relevant features from the existing features and is known as feature selection. The other one is feature extraction where a set of new reduced features is designed based on some transformation function. These techniques may be supervised or unsupervised and it depends on whether they use the output information or not. To analyze the variation and improvement in results, single algorithms are integrated

with these dimensionality reduction techniques which make the hybrid machine learning approach.

## 1.4 THESIS ORGANIZATION

The rest of this thesis is organized as follows. Chapter 2 includes a literature review and a survey of previous work, of different machine learning algorithms for diagnosis of liver disease. Chapter 3 and Chapter 4 describe the scope and present work of study. Present work shows objectives and two research methodologies used in our thesis work to attain the objectives. Chapter 5 describes the experimental work, data analysis and interpretation and performance evaluation of applied algorithms. Chapter 6 concludes with a summary of the thesis contributions and examines further directions for future work. Remaining sections consist appendix and publication part.

# CHAPTER 2
# REVIEW OF LITERATURE

A lot of researchers have implemented various machine learning approaches in order to envisage the presence of liver disease. **A. Gulia et al. [6][2014]** presented various intelligent techniques for liver patient classification. Classification algorithms like (J-48, multi layer perceptron, SVM, random forest and bayesian network) are used for classification of liver patients. All classification algorithms are performed on Indian Liver Patient Dataset (ILPD) from UCI (University of California, Irvine) machine learning repository. Comparative analysis of five different classification algorithms has been done in three phases. In the first phase, simple classification algorithms are applied on the ILPD dataset. In the second phase, a subset of liver patients are selected from whole dataset which consists of significant attributes only and then classification algorithms are applied on the significant subset of attributes . In third phase, the result of classification algorithms is compared. The framework used here is WEKA. The highest accuracy for classification algorithm before applying feature selection is of SVM which is 71.3551. The highest accuracy for classification algorithm after applying feature selection is of random forest which is 71.8696.

**A. Branch et al. [7][2015]** represented different classification models to predict the liver disease is present or not in people based on laboratory parameters. First of all, preprocessing of data is done where steps like normalization of data, determine useless data and determine missing data were performed. Tree decision algorithm, artificial nerve network algorithm, backup arrow machine algorithm, KNN and simple BIZ algorithm were performed to predict liver disease. In this, two datasets were used to assess and compare accuracy of predicting algorithms for liver diseases. One dataset is ILPD and second data set is BUPA liver disorder dataset. Preprocess operation and creating predictive model for liver diseases done by data analysis software RAPID MINER. In ILPD, back up arrow machine model has the high prediction accuracy. In BUPA dataset, back up arrow machine and regression model has the same effectiveness in accuracy of prediction of liver disease.

**B. V. Ramana et al. [8][2011]** discussed various classification techniques for automatic medical diagnosis. In this different classification algorithms are used like

C4.5 decision tree, naive bayes classification (NBC), KNN, SVM and back propagation. In this two liver patient datasets are used, BUPA UCI form machine learning repository and Andhra pradesh liver dataset. Weka data mining open source machine learning software has been used. Classification algorithms' performance has been observed for first four ordered features of AP dataset and then for first five ordered features of AP dataset and soon. After then features of AP Dataset and BUPA UCI are compared and common features are taken. Observed parameters indicate that three common features are very important for accurate diagnosis of liver disease.

**A. S. Aneeshkumar [9][2012]** applied NBC and C4.5 decision tree classification models. Total of 2453 real medical data with 15 attributes were collected from a Public Charitable Hospital in Chennai. The intention of using two algorithms is to identify the improvement of the algorithm for this particular data set. The total data sets have been divided into the ratio of 50-50, 75-25, 90-10 and the accuracy has been evaluated. As a result, the maximum accuracy (99.20%) lies in C4.5 decision tree with 90-10 splitting ratio. But when compare to NBC, it is somewhat lazy, because the average time taken by C4.5 is 0.16 seconds. Naive Bayesian used only 0.03 seconds to achieve the same. Thus, C4.5 decision tree is better than Naive Bayesian. Therefore this application will give more support to such society for their future work and assessments.

**S. Karthik et al. [10][2011]** applied intelligence techniques for diagnosis of liver disease. The classification algorithms are implemented in three phases. In first phase, Artificial Neural Network (ANN) is applied for classification of liver disease followed by second phase where rough set of rules are applied using Learn by Example (LEM) algorithm for classification of liver disease which improves the accuracy. In third phase, in order to identify the types of the liver disease fuzzy rules are applied. Six rules are generated using LEM algorithm which shows an accuracy of 96%. 6% accuracy is improved by LEM algorithm as compared to ANN. Rule extraction is performed to improve the accuracy with correct classified data. The Fuzzy rules are applied to recognize the kind of liver disease. In classification phase, Multi Layer Perceptron (MLP) is employed to differentiate between healthy liver and infected liver. Rule extraction improves the accuracy in the concluding phase.

**E. M. Hashem et al. [11][2014]** presented SVM for analyzing data and pattern recognition. In this, two datasets have been used for classifying liver diseases by using the SVM and then performance is measured on the basis of accuracy, error rate, specificity, sensitivity, and prevalence. One dataset is ILPD dataset and the other is BUPA dataset and then SVM classification algorithm is applied on both the datasets. Ranking of features is done using the ranking algorithm in MATLAB. The specificity of SVM at first 6 ordered features are best for BUPA dataset whereas the sensitivity, prevalence, error rate and accuracy at first 6 ordered features are best for ILPD Liver dataset.

**R. Lin et al. [12][2010].**In this paper, early diagnosis of the liver diseases are done by the intelligent liver diagnosis models. The intelligent liver diagnosis models integrate ANN, Analytic hierarchy process (AHP) and CBR. This integrated model helps us to determine whether the person suffers from the liver disease or not. The type of liver disease is to be found by using CBR. The diagnosis model consists of two steps, the first step is ANN which is implemented in classifying phase and second step is AHP with CBR which is implemented in concluding phase. The CBR has shown the 94.2% similarity of old case with the new case. Thus in the end it can be concluded that hybrid model helps the expert physicians to diagnose the liver disease more accurately and to determine the type of liver disease effectively.

**C. Chuang [13][2011]** developed an auxiliary system in order to diagnose the liver diseases with greater effectiveness and accuracy. CBR is used which is integrated with several data mining classification algorithms like classification and regression trees (CART), back propagation neural network (BPN), logistic regression (LR) and discriminatory analysis (DA) for early diagnosis of the diseases and to increase the classification accuracy. The accuracy, sensitivity, specificity and ROC curve has been shown for every single classification algorithm but CBR added hybrid model shows more accuracy than every single model. BPN-CBR integrated model shows more accuracy than other CBR added integrated models. Thus BPN-CBR outruns other algorithms with a sensitivity of 98%, a specificity of 94% and an accuracy rate of 95% in order to diagnose the liver diseases.

**P. Tamije Selvy et al. [14][2013].**In this paper, for disease diagnosis different classification algorithms are used for increasing the accuracy and the effectiveness. In

data mining where the data is extracted there we perform these algorithms for early diagnosis of the diseases with greater accuracy. The various classification algorithms used here are CBR, KNN, Decision trees, Bayesian Belief Networks, ANN. Of all the algorithms, CBR provides the best results with 92.3% of sensitivity, 90.7% of specificity, and 95.5% of prediction accuracy. Thus these classification algorithms show how the data can be grouped and determined when the new data is available.

**A. Singh et al. [15][2016]** performed diagnosis of liver disease by applying classification methods include LDA, QDA, SVM and feed-forward neural network (FFNN) based approaches. In this, SVM approach has outperformed other classification algorithms on the basis of accuracy rate. Moreover, best predictive model is Least Square SVM (LSSVM) with Gaussian radial basis kernel function machine learning approach. It performs better than linear SVM, polynomial SVM, quadratic SVM and multilayer perceptron SVM in the health examination data.

**X. X. Geng et al. [16][2016]** presented clinical detection of liver cirrhosis by transient elastography. Publication bias, sensitivity analysis and assessment of risk of bias were performed. The area under the receiver operating characteristic (AUROC) curve was 0.931 and the imputed diagnostic odds ratio (DOR) was 26.08. The sensitivity and specificity of transient elastography for detection of liver fibrosis were 81% and 88%. Thus it can be concluded that transient elastography shows good sensitivity, specificity and a high accuracy. It can be utilized as an extra strategy for finding of liver fibrosis and cirrhosis. It supports the use and further development of transient elastography for diagnosing liver fibrosis.

**T. Orczyk et al. [17][2016]** presents a comparative study of various feature selection algorithms. Feature selection algorithm consolidated with chosen machine learning calculations which might be utilized to construct a propelled liver fibrosis determination emotionally supportive network. Characterization calculations incorporate J48 Pruned C4.5 decision tree, IBk KNN classifier, RandomForest forest (ensemble) of random tree classifiers, one rule and decision table classifier. All these are ensemble classifiers. Ensemble of single parameter IBk classifiers using SingleSeparate feature selection method, J48 using CFS feature selection, Random Forest using ReliefF feature selection are the three best classifiers which are showing an accuracy close to 70%. General precision consequences of proposed ensemble

classifiers demonstrates that it is conceivable to construct a solid choice module of a medicinal finding emotionally supportive network utilizing introduced include choice and classification algorithms.

Hepatitis is one of the chronic type of liver disease, in order to diagnose Hepatitis **L. Ozyilmaz et al. [18]** implemented ANN which consists of two standard feed forward network i.e. multilayer perceptron (MLP), radial basis function (RBF) and the other is hybrid network i.e. conic section function neural network (CSFNN). CSFNN has more classification accuracy and is more predictable for Hepatitis diagnosis. Also **I. Campus [19][2011]** determined that supervised models are more accurate as compare to unsupervised models in ANN where feedforward backpropagation neural network (FFNN) , generalized regression neural network (GRNN) belongs to supervised category and self organizing map (SOM) belongs to unsupervised category. To diagnose the Hepatitis B disease **G. S. Uttreshwar [20][2009]** has developed an intelligent model which is based on logic inference integrated with generalized regression neural network and this model shows high predictive classification accuracy which is a convenient choice for accurate diagnosis.

To determine the presence of amount of liver fibrosis in Hepatitis C **R. Stoean et al. [21] [2011]** implemented an evolutionary driven SVM methodology which is very straightforward and extensible and shows better performance as compare to traditional SVM and **A. C. Approach [22][2010]** used different pattern recognition techniques and develop a classification model to determine the degree of liver fibrosis with an accuracy of 93.7%. To diagnose Hepatitis B and C disease **A. G. Floares [23][2009]** has developed an intelligent decision support system based on C5.0 decision tree and figure out an accuracy of 100% of liver biopsy.

**D. Lu et al. [24][2015]** presented multi-determination work division calculation for 3D division of livers, called iterative mesh transformation that misshapes the work of a region of-interest (ROI) in a dynamic manner by cycles between work change and form advancement. Work change deforms the 3D work in light of the distortion exchange display that ventures the ideal work in view of the affine change subjected to an arrangement of limitations of focusing on vertices. In view of this iterative mesh transformation calculation, we built up a semi-computerized plot for division of unhealthy livers with malignancies utilizing as meager as five clients recognized

landmarks. The assessment think about shows that this semi-robotized liver division plan can accomplish precise and solid division comes about with noteworthy diminishment of cooperation time and endeavors when managing infected liver cases.

**C. T. C. Arsene et al. [25][2012]** proposed bayesian neural network (BNN) on primary biliary cirrhosis (PBC) dataset for medical survival analysis which involves two mechanisms : local and global mechanisms and these are used to preserve the numerical stability of PLANN-CR-ARD model. The extent of the review was to foresee the contending events contained in the individual dataset. The numerical outcomes demonstrated that both BNN models could foresee the contending events during the subsequent time and to devise comparable arrangements of prognostic factors. The PLANN-CR-ARD models can be utilized to explore the nonlinear dependency conditions between the anticipated yields and the input information which comprise of the qualities of the PBC patients.

**X. Zhang et al. [26][2015]** proposed texture images for classification of liver fibrosis on computed tomography (CT) or magnetic resonance (MR) images. Diverse sorts of datasets acquired from CT and MR pictures are examined to choose the ideal parameters and elements for the proper order of fibrosis. The general execution computed by the normal whole of most extreme accuracy rate estimation of every one of the 15 elements is 66.83% in CT pictures, while 68.14%, and 71.98% in MR images. Contrasting the precision of characterization and two imaging modalities, the MR pictures havean advantage over CT images with respect to accuracy rate execution.

**D. Li et al. [27][2011]** applied principal component analysis (PCA) to select the suitable subset of features from a given set of features. Then use these transformed data with the optimal subset of features and this will act as the input data to support vector machine (SVM). Datasets related to medicine are generally small and have very large dimensions. When the datasets are small then researchers had applied fuzzy based non-linear transformation to extract related information from the original dataset and this approach shows better results as compare to PCA and Kernel PCA approaches.

**A. Singh et al. [28][2014]** viewed different intelligent techniques and their applications including single ANN, Fuzzy logic, CBR and Genetic algorithm (GA) and data mining techniques etc. Integration of artificial neural network-case-based reasoning (ANN-CBR), artificial immune system-artificial neural network-fuzzy logic (AIS-ANN-FL) is also presented. Different types of liver diseases like liver cirrhosis, liver cancer, liver fibrosis, hepatitis, fatty liver are discussed. The researcher has described that it is totally a specialist's choice to choose with which strategies to continue in view of his experience learning and data he had gotten from this review.

**S. Petrovic et al. [29][2011]** built up a CBR framework for producing dosage anticipates treatment of new tumor patients by catching the experience of oncologists in treating past patients. The proposed CBR framework utilizes an adjusted Dempster–Shafer theory to fuse measurement arranges proposed by the most comparative cases recovered from the case base. Keeping in mind the end goal to mirror the constant learning normal for oncologists, the weights relating to each component utilized as a part of the recovery procedure are refreshed consequently each time in the wake of producing a treatment get ready for another patient.

**V. E. Ekong et al. [30][2015]** represented neuro-fuzzy-CBR driven decision support system in the identification of illness due to depression by utilizing the solutions of past cases. The proposed hybrid framework structure presents a similarity coordinating driven neuro-fuzzy engineering that gives adaptability to doctors measuring the seriousness levels of side effects and manifestations' class. The fuzzified nearby similitudes of manifestations' classes frame the choice factors assessed by the Mamdani-sort ANFIS whose fresh yield speak to the worldwide comparability between the inquiry case and target case. In this work it is additionally proposed CBR and neuro-fluffy mixture system for solving real life problems.

**D. A. Sharaf-el-deen et al. [31][2014]** developed a new integrated approach by combining CBR with the rule-based reasoning (RBR) approach and adaptation process are applied automatically by utilizing adaptation rules. Both adaptation rules and reasoning rules are produced from the case-base. To assess the proposed approach, a model was executed and tested to analyze breast cancer and thyroid maladies. The last outcomes appear that the proposed approach expands the

diagnosing precision of the recovery just CBR frameworks, and gives a solid precision contrasting with the present breast cancer and thyroid diagnosis systems.

**Z. Dong et al. [32][2015]** introduced a medical decision support system based on CBR for identification of plausible tension type headache and to remove the confusion of overlapping between a probable migraine and plausible tension type headache. The consequences of the tests shown a high level of precision in perceiving these two sorts of migraines and a sensational change contrasted with rule based medical decision support system. The exact determination of these sorts of headaches is basic since their medications include distinctive safeguard treatments.

**R. M. Saraiva et al. [33][2015]** presented hybrid CBR and RBR approach to support cancer diagnosis. They utilized manifestations, signs, furthermore, individual data from patients as inputs to demonstrate. To frame specific findings, we utilized guidelines to characterize the input variables' significance as per the patient's qualities. The model's yield exhibits the likelihood of the patient having a kind of tumor. They utilized K-fold cross validation with the information gathered at Napoleao Laureano Healing center. The outcomes demonstrated that their approach is a powerful CBR framework to analyze tumor.

**A. Ghaheri et al. [34][2015]** described that metaheuristic calculations have regularly been utilized as a part of different fields of sciences. In solution, be that as it may, the utilization of these calculations are not known by doctors who may well profit by applying them to take care of complex medical issues. Therefore, they introduced applications of GA in disease diagnosis, prognosis, screening, treatment planning, health care management which helps physicians to analyze the applications of GA in their medical career.

**C.Wu et al. [35][2012]** characterized ultrasonic pictures of liver tissue into the typical liver, hepatoma, and cirrhosis by automatically selecting effective feature subset using GA- based feature selection method. An objective function with consequently decided weightings is outlined for the GA-based feature selection to offer priority to classification rate. The feature cover vector with negligible feature length can be recognized from the arrangement of candidate solutions which accomplishes the highest classification rate.

**Pal et al. [36][2013]** applied continuous GA methodology for developing intelligent computer-aided diagnosis system for heart disease diagnosis. A Consultant doctor's elucidation was utilized to assess the framework's Positive Prediction Value, Negative Prediction Value, Specificity and Sensitivity. The preparatory outcomes appeared promising use for the Medical Multimedia based Clinical Decision Support System as far as right and precise analysis for the unpracticed doctor and in addition reliable and convenient findings, in the investigation of diagnostic protocol, instruction, self-evaluation, and quality control of four noteworthy heart illnesses that were explored.

**E. Sreedevi et al. [37][2015]** developed threshold GA for identification of diabetes disease using Minkowski distance method which is applied on PIMA Indian diabetes dataset. This paper proposed threshold genetic algorithm to analyze the diabetes utilizing Minkowski Distance Method as fitness function to get improved, viable and more exact outcomes. The proposed calculation turns out to be valuable in the part of exactness utilizing PIMA Indian Dataset with different separation strategies. The outcomes got through Minkowski separate technique are nearly superior to other distance methods.

**D. A. Antony et al. [38][2016]** implemented GA to decrease the dimensions of the original dataset which enhance the correctness and performance of classifiers namely J-48, Naïve bayes, and KNN, etc. The proposed GA based feature selection expels the immaterial features and chooses the significant elements from unique dataset with a specific end goal to enhance the execution of the classifiers as far as time to construct the model, lessened dimensions and expanded precision. Thus this proposed strategy is appropriate for the medical application where the algorithms of classification are predefined.

**H. Kahramanli et al. [39][2009]** described that ANN generally shows very high accuracy and the results obtained may be incomprehensible. This fact caused lot of problems in data mining and all these problems was removed by extracting the rules from the ANN. There is a need of improved methods to extract the rules. In this Artificial Immune System (AIS) algorithm was presented to extract the rules. It was implemented on UCI dataset- Hepatitis dataset and Clevelend Heart Disease dataset with accuracy 96.8 % and 96.4 %.

# CHAPTER 3
# RATIONALE AND SCOPE OF THE STUDY

## 3.1 RATIONALE OF THE STUDY

Patients with liver infection have been constantly expanding a result of over the top utilization of liquor, intake of contaminated nourishment, pickles and medications and breathe in of unsafe gasses. It is important to diagnose liver disease early which can increase the patient's survival rate. Expert physicians are required for various examination tests to diagnose the liver disease, but it cannot assure the correct diagnosis. It is very tough job for the expert physicians to recognize the disease from common symptoms. Most of the symptoms are similar to other fever related ailments which result the wrong diagnosis of disease. Accurate classification techniques are required for automatic identification of disease samples. This study accordingly employs various individual and integrated machine learning algorithms on distinct liver disease datasets to evaluate the diagnostic performances in terms of different parameters.

## 3.2 SCOPE OF THE STUDY

The scope of the study will focus on the correct diagnosis of liver patients as early as possible. The scope of this research is as follows:

- To use classification approaches that help successful early liver diagnosis and treatment. As patients are not ready to spend more time in the hospital and they need to be cure immediately.

- To implement classification approaches as automatic or real time classification tools which may useful for experts to identify the chances of disease and conscious prescription of further medical examinations and treatment. The reason is that recent days every hospital is rich with large amount of medical records but without effective analysis in that. Most patients admitted in hospitals with multiple diseases and most of the diseases have multiple symptoms. In such cases it is very complex to diagnose each disease and treats separately so they may get

treated for major disease and the upcoming diseases may not be identified or neglected.

- This research will help medical practitioners to demonstrate awareness of evidence based treatment and therefore this application will give more support to such society for their future work and assessments.

- Moreover, this also helps number of trainees for performing their research work and practices.

# CHAPTER 4
# PRESENT WORK

## 4.1 PROBLEM FORMULATION

The research problem is a crucial part of any research activity. If nature of the problem is clear that it is very easy to solve the problem.

Liver diseases represent a major health burden worldwide. It is rising day by day and is not discovered easily in its initial stage as liver can work properly even when it is partially damaged. Machine learning algorithms have been extensively used by doctors for diagnosing liver disease so as to enhance efficiency of medical diagnosis. The problem of the study is to determine the applications of machine learning algorithms in liver disease diagnosis on distinct liver datasets. This aims to determine the performance and results of various single and integrated machine learning algorithms on the basis of different parameters. This problem materializes into following sub-questions:

- To predict whether one particular single or integrated machine learning algorithm shows best result for variant liver datasets or not.

- If the best result is not same for all datasets then need to analyze why there are variations in their results.

- How we can make sure that one particular algorithm can't be perfect for all datasets and performance of an algorithm depends only on the dataset type and structure.

## 4.2 OBJECTIVES OF STUDY

Any task without sound objectives is like tree without roots. In the same way during any research undertaken, first objectives of research study are determined and then next steps are taken in order to proceed further. A research study may have many objectives but all these objectives revolve around one major objective which is the focus of study.

Liver diseases represent a major health burden worldwide. It is rising day by day and is not discovered easily in its initial stage as liver can work properly even when it is partially damaged. Machine learning algorithms have been extensively used by doctors for diagnosing liver disease in order to enhance efficiency of medical diagnosis. The main objective of the study is to analyze the applications of machine learning algorithms to avoid delaying the medical treatment, for correct diagnosis of the liver disease and to reduce the false diagnosis given to sick people. The following objectives have been set forth. They are:

1. To implement various machine learning algorithms on distinct liver disease datasets for evaluating the diagnostic performance.
2. To integrate dimensionality reduction and optimization techniques with the selected classification algorithms for analyzing the variation in results.
3. To find the best classification models for liver diseases.
4. To analyze the merits and demerits of applied techniques.

## 4.3 RESEARCH METHODOLOGY

### 4.3.1 Proposed Methodology I

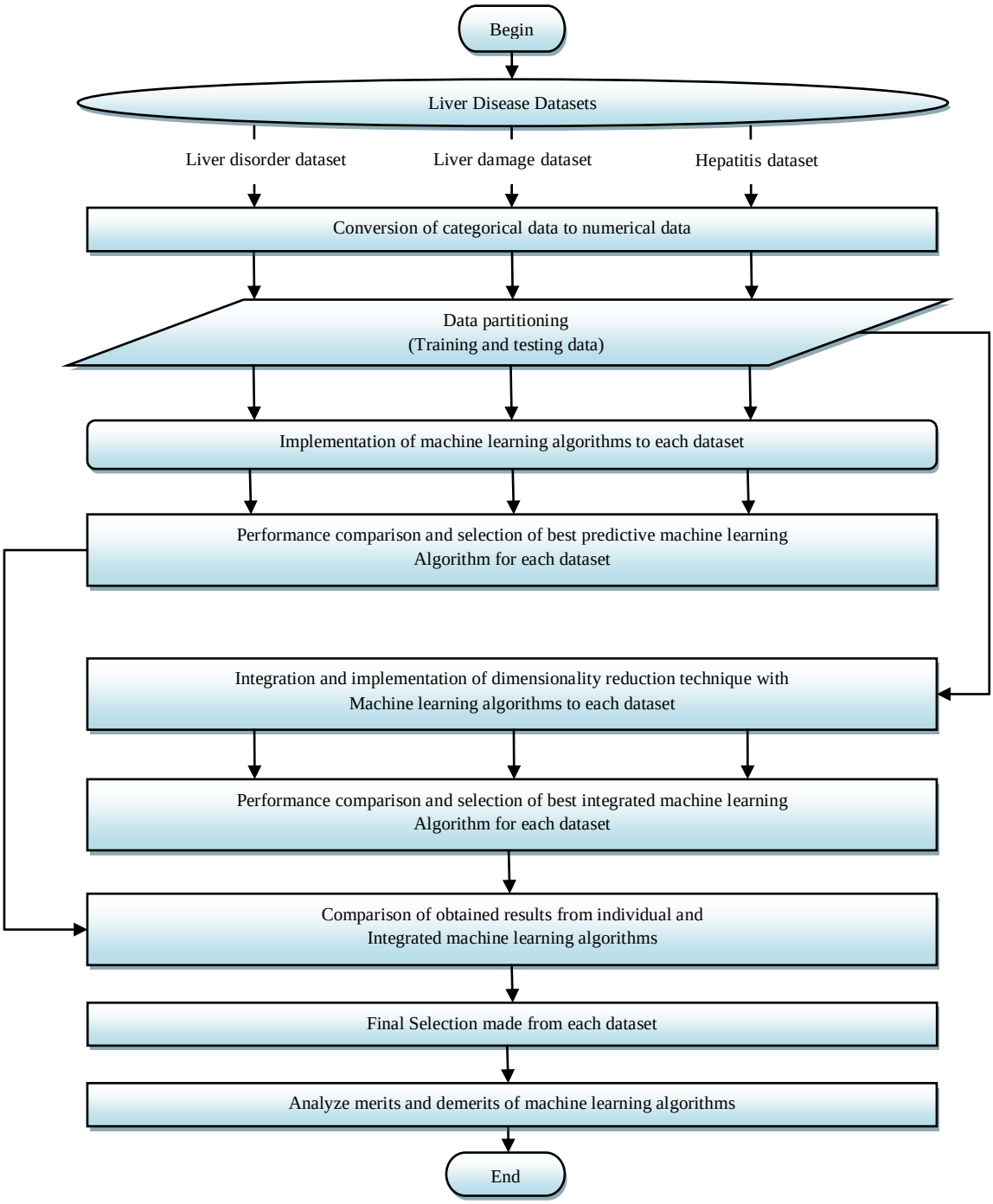


**Figure 4.1** The Overall System design for the diagnosis of liver diseases

#### *4.3.1.1 Machine Learning Algorithms*

Classification algorithms are widely used in various medical applications. Data classification consists two phases – one is training phase and the other is testing phase. In training phase, a classifier is built by the classification algorithm with only training set of tuples and in testing phase, performance of a classifier is evaluated with testing set of tuples. In this study different classification algorithms are considered which include linear discriminant analysis (LDA), diagonal linear discriminant analysis (DLDA), quadratic discriminant analysis (QDA), diagonal quadratic discriminant analysis (DQDA), support vector machine (SVM) and K nearest neighbor (KNN) Algorithm. All these algorithms are applied on three different liver datasets to analyze their performances in terms of different parameters. Afterwards, these algorithms are integrated with the dimensionality reduction approach which is principal component analysis (PCA) in order to decrease the dimensions and to analyze the variation in results. All these classification algorithms and the PCA approach are as follows:

**1. Linear Discriminant Analysis and Diagonal Linear Discriminant Analysis**

In discriminant analysis, there are two variables: dependent and independent variable. The dependent variable (Y) represents the group and any dataset features that describe the group are independent variables (X). The variable which is dependent is known as the category variable also called nominal scale and the variable which is independent can be of any measurement scale like nominal, ordinal, interval, ratio scale. Linear discriminant function can be used if the groups are linearly separable. In this, linear separable means that linear combination of features can separate the groups. If there are only two features then lines will become a separator between the objects group, if there are three features then plane will become separator and if more than three features are present then the separator will become hyper-plane.

LDA is known for its easiness where both time and space complexities are Θ (d). The weighted sum of input attributes ($y_k$) represents the final output. The influence of $y_k$ is represented by magnitude of weight $w_k$ and sign demonstrates either the effect is positive or negative. LDA also shows good accuracy when optimal discriminant is linear and classes are Gaussian with shared covariance matrix. But it

can also be used even when this assumption does not hold and we can easily calculate the model parameters even without making any assumptions on the class densities.

DLDA is somewhat similar to LDA, the only difference is that in this DLDA we use the diagonal of 'linear' covariance matrix means we can take only the diagonal elements of covariance matrix and even we can use the pseudoinverse if it is necessary where pseudoinverse is a inverse of matrix and it is like a object which may be defined for complex matrix but it is not significantly a square. More than one pseudoinverse may also be possible for a given covariance matrix.

## 2. Quadratic Discriminant Analysis and Diagonal Quadratic Discriminant Analysis

QDA is almost similar to LDA. The difference is, covariance matrix and mean is different for each class. For each class m, m =1, 2, ..., M there is need to estimate the covariance matrix $\Sigma_m$ separately. QDA has the ability to fit the data more suitably as compare to LDA because of its more flexibility for the covariance matrix but there is need to estimate more parameters. The number of parameters in QDA increases significantly. As, with QDA, there will be different covariance matrix for each class. The problem can arise if there are more classes and not so many sample points.

DQDA is somewhat similar to QDA. The only difference is that, in DQDA the diagonal of the 'quadratic' covariance matrix is used. Only the diagonal elements of covariance matrix are taken. Here pseudoinverse can be used if it is necessary, where pseudoinverse is an inverse of matrix. It is like an object which may be defined for complex matrix but it is not significantly a square. More than one pseudoinverse may also be possible for a given covariance matrix.

## 3. Support Vector Machine

SVM is a supervised learning method used for both classification and regression. It has very high generalization performance and there is no requirement to add a prior knowledge, even when it has very high input space dimension and this makes it a very good quality classifier. The main intend of the SVM classifier is to discriminate between members of two classes in the training data by finding best classification function. SVM is a generalized linear classification method and it simultaneously maximizes the geometric margin and minimizes the classification error. In n

dimensional space, viewing input data as two sets of vectors, a separating hyper-plane will be constructed by SVM which maximizes the margin between two data sets. Two parallel hyper-planes are constructed in order to calculate the margin, one on each side of separating hyper-plane [40]. The largest distance to the neighboring data points of both the classes helps to achieve good separation and if the margin is large then the generalization error will become less. Thus, the support vectors and margins help to find the hyper-planes.

Here the data points are considered in the form of:

$\{(z_1, x_1), (z_2, x_2), (z_3, x_3), (z_4, x_4) \dots \dots \dots \dots, (z_i, x_i)\}$

Here $z_i = 1/-1$, is a constant which donates the class to which $z_i$ belongs where i is the number of samples. $z_i$ represents m-dimensional real vector. With the help of separating hyperplane we can easily view the training data, which is

$$u.z + c = 0 \quad (1)$$

where u is m-dimensional vector and c is scalar.

There is a vector u which is perpendicular to the dividing hyperplane and the scalar parameter c helps to increase the margin. In case of absence of c the hyper-plane is passed through the origin by force. In order to maximize the margin, parallel hyper-planes are required. These parallel hyper-planes are described by the equation.

$u.z + c = 1$

$u.z + c = -1$

If training data is separated linearly, we can choose the parallel hyper-planes as there are no points among them and can also give a try to maximize the distance or margin. Here by geometrically we can find the distance between the hyperplane which is $2/|u|$. This is the reason why we want to lessen the $|u|$. The equation is:

$u.z_j - c \geq 1 \quad or \quad u.z_j - c \leq -1$

The following equation gives the distance between the point $z$ and a hyperplane $(\alpha, \alpha_0)$:

$$distance = \frac{|\alpha_0 + \alpha^C z|}{\|\alpha\|} \quad (2)$$

**Algorithm: Pseudo-code for SVM**

1) Propose positive lagrange multipliers, one for each of the inequality constraints. This offers lagrangian:

$$L_m = \frac{1}{2} \|u\|^2 - \sum_{j=1}^{i} \beta_j \, x_j (z_j . u - c) + \sum_{j=1}^{i} \beta_j$$

2) Minimize $L_p$ with relevance to u,c. This is a convex quadratic programming problem.
3) In the solution, those points for which $\beta_j > 0$ are called "support vectors".

Model selection of SVM is also a difficult approach. SVM has shown a good performance in data classification recently. Tuning of several parameters is an effective approach which affects the generalization error and this acts as the model selection procedure. In case of linear SVM there is a need to tune the cost parameter C. But linear SVM is generally applied to linearly separable problems. In cross validation we can use the grid search method to find the paramount parameter set. Then we obtain the classifier after applying this parameter set to the training dataset and this classifier is used to classify the testing dataset to obtain the generalization accuracy.

## 4. K-Nearest Neighbor Classification Algorithm

KNN algorithm is one of the algorithms which are very easy to understand and should be one of the choices in case of classification study and also it works incogitable well in practice. KNN is also known as non parametric lazy algorithm. Here non parametric means no assumption can be made by this algorithm on the underlying data distribution and lazy algorithm means in order to do any generalization, it does not use any training data points. Thus it means there is no explicit training phase. We can say that training phase is little bit fast. We need all the training data in testing phase. Here training phase is minimal but testing phase is costly and the cost is both in terms of time and memory. There is one assumption in KNN that data is in feature space. The data can be in scalar form or in multidimensional vector form . Here the training data can be expressed by set of vectors and with each vector a class label is associated. For positive and negative classes it will be either + or -. Here the value of

k defines the number of neighbors that represents the significance of classification and neighbors are selected on the basis of distance metric.

**Algorithm: The KNN Algorithm for approximating a discrete-valued function $l: P^i \rightarrow Y$.**

Training algorithm:

- For each training example *(z, l ( z ) )* , upload the instance to the list training_examples

Classification algorithm:

- Given a query example $z_r$ to be labeled,
    - Let $z_1, \ldots \ldots \ldots z_k$ represent the *k* instances from *training_examples* that are nearest to $z_r$
    - Return

$$\hat{l}\,(z_r) \leftarrow \underset{y \epsilon Y}{\operatorname{argmax}} \sum_{j=1}^{k} \delta\ (y, l(z_j))$$

where $\delta(b, c) = 1$ if $b = c$ and where $\delta(b, c) = 0$ otherwise.

This algorithm returns the value of $\hat{l}\,(z_r)$ which describes the most common value of $l$ from the k training examples which are nearest to $z_r$. If we choose the value of k=1 then we can assign the 1-nearest neighbor algorithm to $\hat{l}\,(z_r)$ the value of $l(z_j)$ where the training instance nearest to $z_r$ is $z_j$. If value of k is large then the most common value among the k nearest training examples is assigned by the algorithm. We can easily adapt KNN algorithm in order to approximate continuous valued target functions. Rather than calculating their most common value, we have an algorithm to calculate the mean value of k nearest training examples. To approximate the real value target function $l: P^i \rightarrow P$ we have to replace the final line of the above algorithm by the line

$$\hat{l}\,(z_r) \leftarrow \frac{\sum_{j=1}^{i} l(z_j)}{k} \qquad (3)$$

KNN algorithm is used mostly when we don't have any prior knowledge about the distribution of data and for classification study it can be one of the best choices. We can use the same method in case of regression also where we have to assign the property values for object to observe the average of the values of k nearest neighbors.

**5. Principal Component Analysis**

PCA is a popular dimensionality reduction technique which transforms the large number of correlated variables into less number of variables using various mathematical principles and these variables are known as principal components. To reduce the dimensionality of large datasets, PCA generally uses vector space transformation method where the original dataset with large number of variables are rehabilitated into small number of variables called principal components. Thus this reduced dimension dataset helps the user to analyze the patterns, outliers in data which is very difficult to perform without performing PCA [41].

If there are p variables then p principal components can be formed where the first principal component is formed by linear combination of variables that has the largest variance and also other succeeding component can be formed by the linear combination of variables with largest possible variance but the components are uncorrelated with the previous components. We can calculate the principal component by taking the linear combination of an eigen vector of correlation matrix with the variables. The eigen values present in the eigen vector actually represent the variance of each component.

**Algorithm: PCA (E, O)**

1) $\mu \leftarrow mean\ (Y)$ // compute data mean for centring
2) $E \leftarrow (Y - \mu 1^C)^C\ (Y - \mu 1^C)$ // compute covariance
3) $\{\alpha_o, \mu_o\} \leftarrow$ top O eigen values/eigen vectors of E
4) Return $(Y - \mu 1)\ W$ // project data using W.

PCA is a useful statistical technique which is found in various application fields like image compression and face recognition and is a common technique for finding patterns in high dimensional data. But it is necessary to note that PCA is very sensitive to the scaling of data and to obtain the best results there is no consensus as to how to best scale the data.

*Benefits of Dimension Reduction*

- It helps in compressing the data.

- It reduces the storage space required.
- The time required for performing the same computations is less as less dimensions leads to less computing.
- Fewer dimensions can allow the usage of algorithms unfit for a large number of dimensions.
- It maintains the multi-collinearity which improves the model performance.
- It removes redundant features.

For high dimensional datasets, dimension reduction is usually performed prior to applying machine learning algorithms in order to avoid the effect of curse of dimensionality.

### 4.3.2 Proposed Methodology II

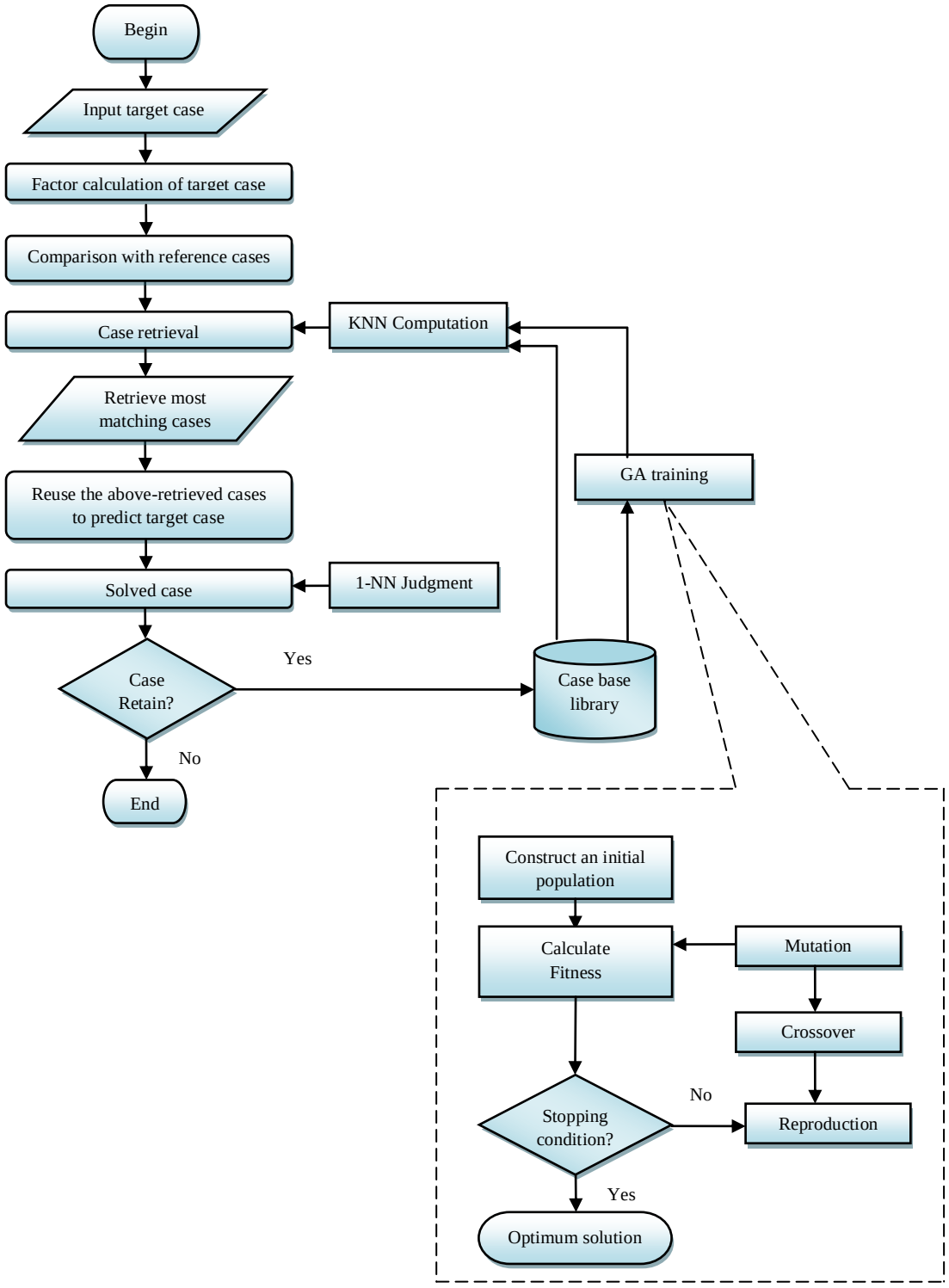


**Figure 4.2** A CBR-GA based integrated model for diagnosis of liver diseases.

#### *4.3.2.1 Integrated GA and CBR model*

To develop an integrated framework of GA and CBR model there are some steps need to perform as demonstrated in Figure 5.2. In the first step, target case is inserted into the case base related to liver disease dataset. Then target case matches with the previous cases of liver disease dataset to find the proper solution to the problem. Then determine the factors related to the target case. When the target case is added then compare the target case with reference cases from the historical case-base to find the matching case to given problem. The comparison is made by similarity rule to predict whether the person suffers from liver disease or not.

In the case retrieval step, the most matching cases from the previous cases can be retrieved by using KNN computation. KNN finds the k most matching cases from the previous case base to find the solution to the new problem. For example, if the results of 5 similar cases are showing diseased persons then the result of the target case would be a diseased person for sure. In order to calculate the optimal weight of each factor, GA training is applied. The solution of retrieved case is reused directly to solve the problem of the target case. After then, put the revised case in the historical case-base for the discovery of knowledge in future and this procedure is known as retention. The termination condition occurs if we do not want to put the reused or revised case in the historical case base library.

#### *4.3.2.2 Case Based Reasoning (CBR)*

CBR is an approach of constructing knowledge based frameworks which are used to solve different problems. It adapts the prior solutions in order to find the solutions to target problems. In CBR terms, a case normally indicates a problem situation. A case represents features in matrix form. The basic principle behind the CBR methodology is that same type of solutions can be retrieved if the problem is of the same type. CBR does not require in-depth analysis to solve the new target problems which make it differ from other artificial intelligent techniques. The reasoning procedure of CBR is explained in following steps:

*Step 1: Indices and weights assignment*

Input the training examples of real world liver disease dataset in medical record interface. In light of the expert conclusions from the gastroenterologist, a case index is worked to fuse different attributes. The weight is assigned to each feature to analyze the importance of feature which helps to measure the similarity. If the weight of a feature is high, the significance of that feature is also considered to be high. Therefore, assignment of proper weights to the features is an important task [12].

*Step 2: Case Retrieval*

In CBR, the purpose of this step is to retrieve the matching cases to the problems as fast as conceivable. Problem description is an input to the task of retrieval and output represents the cases that most closely match the target problem. Relevant cases are found according to the similarity in the features which can be measured by comparing the target with the other cases in the case base. Similarity measure totally depends on the weights allotted to attributes. Similarity calculation can be done for each quantitative or qualitative feature.

Qualitative similarity: The similarity between corresponding classes is defined by $\mathrm{sim}(y,z)$ as described in Eq. (4). This is known as qualitative similarity and it mainly exists between categorical features.

$$\mathrm{sim}(y,z)=\begin{cases}1 \text{ if } y=z, & \text{similarity in the features belong to same class.}\\ 0 \text{ if } y\neq z, & \text{similarity in the features belong to different classes.}\end{cases} \quad (4)$$

Quantitative similarity: It is a similarity measure of real or integer number with lower and upper bounds. The similarity measure value lies between 0 and 1.

$$\text{if} \quad s=t \text{ then similarity} = 1$$
$$\text{else} \quad \mathrm{sim}(s,t) = 1.0 - \left[\frac{|s-t|}{(up-lw)}\right] \quad (5)$$

where s is the query value, t represents the case value, up is the upper bound of number and lw is the lower bound of number.

Overall similarity: Similarity of both quantitative and qualitative features helps to calculate the overall similarity**.** The similarity score which is defined in Eq. (6)

helps to retrieve the matching case from the case base library using the nearest neighbor computation.

$$\text{similarity score} = \frac{\sum_{j=1}^{k} W_j \, \text{sim}\left(f_j^I, f_j^R\right)}{\sum_{j=1}^{k} W_j} \tag{6}$$

where $W_j$ is weight of index j and $\sum_{j=1}^{k} W_j = 1$. sim is similarity function for primitives. $f_j^I$ and $f_j^R$ are values of input and retrieved cases for feature $f_i$.Retrieval task from the case base can be defined as the selection of small number of cases with the highest similarity to the query. The goal of matching task is to find the cases that are similar to the target problem and can find the solution to that problem. If this step has been executed completely then go to the step 3 else perform step 4.

*Step 3: Case Reuse*

Reuse the case is the third step. After finding matching cases to the current case, the framework needs to reason as per the retrieved cases to locate a sensible and exact answer for the problem. Reusing the case can be done in two ways: the first one is considering diagnosis and classification application, where the solution of a retrieved case is directly used as the solution of target case. The other way represents frameworks which depend on adaptation strategies. In this, the result of retrieved cases is not directly used as the solution to a new problem.

*Step 4: Case Adaptation*

This step represents the tasks where we do not have possible solutions in the case base to solve a particular problem. Retrieve the matching cases, then find the similar solutions and use the difference between the target case and retrieved case in order to alter the retrieved solution for the target problem. This is called the revising and optimization of the case. After performing this step go to step 5.

*Step 5: Retain Case*

Reuse the cases and put the revised case in the liver disease case base for discovery of knowledge in future. This process is known as retention. Therefore, the solution of different problems can be used for further future problem solving mechanisms. However, sometimes this step leads to uncontrolled growth of the case

base due to continuous retention of different cases which decreases the speed and ultimately reduces the performance of the system. Maintenance of the case base is necessary for these problems.

#### *4.3.2.3 Genetic Algorithm (GA)*

Evolutionary Computation (EC) consist a number of techniques and approaches based on natural selection. GA is one of the algorithms that are generally applied to problems based on classifier systems and evolutionary strategies. It is an optimization technique which is inspired by natural genetics and natural selection. GA is an effective approach for finding solutions to different problems using optimization techniques. GA continuously tries to find various possible solutions using different genetic operators like selection, crossover, and mutation. CBR principally concentrates on the most proficient method to depict and retrieve cases. Weight is generated for each factor by using CBR and the best possible weight for each factor in CBR is discovered by GA [34]. The steps for finding the best combination are as follows:

*Step 1: Encoding in GA*

The first step to solve a problem with GA is encoding of chromosomes. The process of encoding completely depends on the problem. Binary encoding is the most frequent method of encoding. Every chromosome is represented by a string of bits i.e. 0 or 1 and its combination is used in weight assignment to each factor. This encoding is regularly not normal for some problems as some remedies must be required after crossover or mutation. Each factor influencing liver disease is assigned a weight with the combination of eight binary numbers.

*Step 2: Initialize the population*

The initial population is generated after the encoding step of GA. Every individual of population encodes a conceivable answer for a problem. The size of the population depends on the size of problems search space and computational time taken by it to evaluate each individual.

*Step 3: Evaluation of Chromosomes*

After constructing the initial population, evaluation of each individual is done and a fitness value is assigned in accordance with the fitness function. The

performance of every chromosome is evaluated by fitness function and it determines how closely it matches to the solution. The comparison is made between the fitness values of a new chromosome and current chromosome. If fitness value for new chromosome is high then it will be reserved for new offspring.

In this, fitness value is calculated for each case of liver disease laboratory factors to predict whether the person suffers from liver disease or not. Fitness value for the training cases is as follows:

Objective function:

$$\max M(A) = \sum_{j=1}^{n} R_j \quad (7)$$

s.t.

$R_j = 1, \text{if } L_j = V_j$

$R_j = 0, \text{if } L_j \neq V_j$

where $M(A)$ is an objective function to predict liver disease for a set of training cases A. $L_j$ refers the predicted result of case j in training cases: If case j represents a person with liver disease then $L_j = 1$, otherwise the person is free from liver disease and $L_j = 0$. $V_j$ refers the actual result of case j in training cases: If case j represents a person with liver disease then $V_j = 1$, otherwise the person is free from liver disease and $V_j = 0$. $R_j$ defines the prediction and actual result comparison for case j, if the result is similar then $R_j = 1$ else $R_j = 0$.

Calculation of $L_j$:

Let us consider,

Reference case set $C = \{c_1, c_2, \dots\dots\dots, c_m\}$, $p = 1,2,3, \dots, m$

Training case set $A = \{a_1, a_2, \dots\dots\dots, a_n\}$, $j = 1,2,3, \dots, n$

$$L_j = Z(c_p), \text{if } S_{jp} = \min_{p}[D(c_p, a_j)]$$

and

$$D(c_p, a_j) = \sqrt{\sum_{z=1}^{k} w_z \left(f_{pz} - f_{jz}\right)^2}, \qquad z = 1,2, \dots, k$$

where k is the total number of factors, $S_{jp}$ is the degree of similarity between case j and case p of the training set A and reference set C. D represents the aggregation of separations between every weighted component of training and reference case set. $Z(c_p)$ refers the consequence of case p of reference set C which is most like to case j of training set A, $c_p$ represents the diseased person then $Z(c_p) = 1$, otherwise $Z(c_p) = 0$. $w_z$ is the weight of variable z in reference cases. $f_{pz}$ represents the Estimation of variable z of case p in reference cases. $f_{jz}$ is the estimation of variable z of case j in training cases.

*Step 4: Fitness value computation*

The performance of every chromosome is evaluated by fitness function and it determines how closely it matches to the solution. The concept of fitness is related to the fact that highest the fitness value, highest it tends to propagate to the next generation. Fitness function finds the total number of accurate classifications for the entire dataset. The objective function for the problem of prediction of liver disease patients described in this research is to find the accuracy value. Thus, for a set of training cases an objective function acts as a fitness function.

$$\mathrm{fit}(A) = M(A) = \sum_{j=1}^{n} R_j \tag{8}$$

*Step 5: Reproduction/Selection*

Reproduction or selection is the first operator which is applied on the population. This operator uses the proportional selection of population. This step involves the roulette wheel selection method which means the probability of each individual is equivalent to the fitness function $\mathrm{fit}(y)$ in every generation to the aggregate fitness value $\sum \mathrm{fit}(y)$ of whole population individuals. This represents if the fitness value is higher, then the probability of being selected into next generation is also higher. Each chromosome y will be chosen to reproduce with probability $\mathrm{pr}(y)$ is defined as:

$$\mathrm{pr}(y) = \frac{\mathrm{fit}(y)}{\sum \mathrm{fit}(y)} \tag{9}$$

*Step 6: Crossover*

Crossover is the second operator applied on the population where two individuals are produced by selecting two parents. Crossover probability helps to perform two-point crossover operation in order to construct two new chromosomes.

*Step 7: Mutation*

The random substitution method is used in this study to perform mutation operation. It is the third operator which is applied to the population. In this step, the chromosome which is to be mutated is replaced by new randomly generated chromosome with the low probability value.

*Step 8: Update*

The old individuals with poorer fitness value are replaced by new individuals after evaluating new individuals from crossover and mutation operations. This step is known as Replacement or Update step.

*Step 9: Termination condition*

The process of GA has repeated until the number of genetic cycles reached the maximum number of predefined genetic cycle number.

# CHAPTER 5
# RESULTS AND DISCUSSION

## 5.1 EXPERIMENTAL WORK

### 5.1.1 Tools description/ Simulation work

In this thesis work, Matlab Tool and CBR Applet Viewer tool is used in order to implement different techniques on different datasets which are defined as follows:

#### *5.1.1.1 Matlab Tool*

Matlab stands for MATrix LABoratory is defined as the language providing high performance in technical computing and it is easy to use in environment where everything is in mathematical equations. MATLAB is generally utilized as a part of all ranges of connected arithmetic, in instruction and research at colleges, and in the business. MATLAB has effective realistic instruments and can create pleasant pictures in both 2D and 3D. It is additionally a programming language, and is one of the simplest programming languages for composing scientific projects. MATLAB likewise has some tool compartments valuable for, picture preparing, improvement, signal processing and so forth. It is used in different areas like:

- Engineering and scientific explorations
- Modeling and simulation
- Data Acquisition
- Graphical User Interface building
- Algorithm development

Matlab tool consists of the following parts:

Mathematical Function Library: It constitutes vast collection of computing algorithms like matrix inverse, fast fourier transforms and Bessel functions.

Language: It is high level matrix/array language with control flow statements, object oriented programming features and data structures etc.

External Interfaces: It allows writing C and FORTRAN programs that interact with Matlab.

Graphics: It provides two and three dimensional data visualization, image processing and animation.

Desktop Tools and Development environment: It provides the tools like command window, editor and debugger, code analyzer etc.

#### *5.1.1.2 CBR Applet Tool*

Case-Based Reasoning is one of best connected AI innovations of latest years. Business and mechanical applications can be produced quickly and existing corporate databases can be utilized as learning sources. Helpdesks and indicative frameworks are the most well-known applications.

CBR shell is a generic tool for case based reasoning. Classification is performed by case comparison. Leave-one-out is an evaluation method to measure the performance of an algorithm. It is a cross validation approach where training is performed on all data except for one point and prediction is made for that point. The matching cases to the new case can be found by using K-NN method. This CBR shell has following different features.

- Genetic algorithm weight learning
- Multiple diagnostic algorithms
- Cross-platform implementation
- K nearest neighbor and threshold retrieval

### 5.1.2 Techniques and Methodologies used

In this research work, two methodologies are used. In first methodology, a system is designed to diagnose the presence of liver disease. Six different techniques or algorithms are used which includes LDA, DLDA, QDA, DQDA, SVM and KNN. These algorithms are applied on three liver disease datasets: BUPA liver disorder dataset, liver damage dataset and hepatitis dataset to evaluate their performances. In order to analyze the variation in results, all these algorithms are integrated with dimensionality reduction method i.e. PCA which is an integrated approach. Then comparison is done between different algorithms to find the best algorithm.

In second methodology, CBR technique is applied on the above said three liver disease datasets. In order to optimize the results, CBR is integrated with GA. An

integrated GA-CBR model is proposed in this study to compare the effectiveness of this model with CBR in the prediction of liver disease. This model finds the impact of optimization algorithm on prediction results of CBR in liver disease. Optimization of weights of features and selection of appropriate instances for CBR is done simultaneously in this proposed model.

## 5.2 DATA ANALYSIS AND INTERPRETATION

In this study, we used three datasets related to liver disease where one dataset is Liver damage dataset which has been taken from AstraZeneca healthcare foundation and the other two are taken from UCI (University of California, Irvine) machine learning repository which includes BUPA liver disorder dataset, Hepatitis dataset. The liver damage dataset consists of five attributes ALP alkaline phosphate, ALT alanine aminotransferase, AST aspartate aminotransferase, TBL Total bilirubin, a dose which is a factor with levels A, B, C and D and 606 observations. The BUPA liver disorder dataset includes the data of 345 patients with 6 independent variables and 1 dependent variable. The independent variables are mcv, alkphos, sgpt, sgot, gammagt, drinks. The dependent variable is selector field with two categories – 1 or 2 on the basis of presence or absence of liver disease. The Hepatitis dataset has 155 instances with 19 independent variables and 1 dependent variable. The independent variables are age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology. On the other hand, the dependent variable represents a class with two categories 1 (DIE) or 2 (LIVE).

Table 5.1, Table 5.2 and Table 5.3 describes the name, representation and range values of BUPA liver disorder dataset, Liver damage dataset and Hepatitis dataset.

**Table 5.1** Feature details of BUPA liver disorder dataset.

| Variable Name | Meaning | Represented as | Intervals |
|---|---|---|---|
| **Mcv** | Mean corpuscular volume | Integer | 65-103 |
| **Alkphos** | Alkaline phosphotase | Integer | 23-138 |
| **Sgpt** | Alamine aminotransferase | Integer | 4-155 |
| **Sgot** | Aspartate aminotransferase | Integer | 5-82 |
| **Gammagt** | Gamma-glutamyl transpeptidase | Integer | 5-297 |
| **Drinks** | Number of half-pint equivalents of alcoholic beverages drunk per day | Real | 0-20 |

**Table 5.2** Feature details of Liver damage dataset.

| Variable Name | Meaning | Represented as | Intervals |
|---|---|---|---|
| **ALP** | Alkaline phosphatase | Integer | 15-129 |
| **ALT** | Alanine aminotransferase | Integer | 4-198 |
| **AST** | Aspartate aminotransferase | Integer | 5-104 |
| **TBL** | Total bilirubin at baseline | Real | 2.736-27.531 |

**Table 5.3** Feature details of Hepatitis dataset.

| Variable Name | Meaning | Represented as | Intervals |
|---|---|---|---|
| **Age** | Lifetime in years | Integer | 7-78 |
| **Sex** | Gender | Integer | Male, Female |
| **Steroid** | An organic compound with four rings | Integer | No, Yes |
| **Antivirals** | An agent that kills a virus | Integer | No, Yes |
| **Fatigue** | Tiredness | Integer | No, Yes |
| **Malaise** | Discomfort or illness | Integer | No, Yes |
| **Anorexia** | Loss of appetite | Integer | No, Yes |
| **Liver Big** | Enlarged or swollen liver | Integer | No, Yes |
| **Liver Firm** | Abnormalities of liver | Integer | No, Yes |
| **Spleen Palpable** | Enlargement of spleen | Integer | No, Yes |
| **Spiders** | Thin vessels form a web like shape | Integer | No, Yes |
| **Ascites** | abnormal accumulation fluid in the abdominal cavity | Integer | No, Yes |
| **Varices** | Enlarged veins | Integer | No, Yes |
| **Bilirubin** | Yellow pigment in blood | Real | 0.3-8 |
| **Alk Phosphate** | Protein found in body tissues | Integer | 26-295 |
| **Sgot** | Serum glutamic oxaloacetic transaminase | Integer | 14-648 |
| **Albumin** | Protein found in blood | Real | 2.1-6.4 |
| **Protime** | Prothrombin time (PT) is a blood test | Integer | 0-100 |
| **Histology** | The study of the microscopic structure of tissues | Integer | No, Yes |

## 5.3 PERFORMANCE EVALUATION

### 5.3.1 Performance Evaluation Parameters

Performance of applied algorithms are measured by using different parameters like accuracy, specificity, sensitivity, negative prediction value (NPV) and positive prediction value (PPV). Some basic definitions are required for the explanation of these parameters.

**Definitions:**

Patient: positive in case of disease

Healthy: negative in case of disease

True negative (TN) = the number of cases effectively distinguished as healthy.

False negative (FN) = the number of cases erroneously distinguished as healthy.

True positive (TP) = the number of cases effectively distinguished as patient.

False positive (FP) = the number of cases erroneously distinguished as patient.

1. Accuracy: The accuracy of a test is its capacity to separate the patient and sound cases accurately. We should calculate true positive and true negative in proportion for all evaluated cases. The mathematical formula for accuracy is as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

2. Specificity: The specificity of a test is its capacity to decide the healthy cases accurately. To gauge it, we ought to compute the proportion of true negative in healthy cases. Mathematically, this can be expressed as:

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

3. Sensitivity: The sensitivity of a test is its capacity to decide the patient cases accurately. To gauge it, we ought to figure the proportion of true positive in patient cases. The mathematical formula for sensitivity is expressed as follows:

$$Sensitivity = \frac{TP}{TP+FN} \quad (12)$$

4. Positive Prediction Value (PPV): PPV is the proportion of positive results in diagnostic tests that are true positive results respectively. The mathematical formula of PPV is expressed as follows:

$$PPV = \frac{TP}{TP+FP} \quad (13)$$

5. Negative Prediction Value (NPV): NPV is the proportion of negative results in diagnostic tests that are true negative results respectively. The mathematical formula of NPV is expressed as follows:

$$NPV = \frac{TN}{TN+FN} \quad (14)$$

### 5.3.2 Experimental Results of Methodology 1

In each compared model or algorithm data is divided into two sets – training set and testing set by using hold out cross validation approach where data is divided in the ratio of 80:20 i.e. 80% of data is used to train the model and 20% of data is used to test the model. In training phase, a classifier is trained whereas in testing phase, the testing of classifier is done to analyze performance. Both the training and testing data remained same for all compared methods. The performances of all algorithms are evaluated in terms of accuracy, sensitivity, specificity, positive prediction value (PPV) and negative prediction value (NPV). The software we espoused in this study is MATLAB R2013a with its powerful bioinformatics and statistics for machine learning toolbox. The experimental results of all the single method models LDA, DLDA, QDA, DQDA, SVM, KNN and all the PCA-added integrated models are presented separately in the following section.

#### *5.3.2.1 Performance Evaluation of Single Models on Liver Datasets*

Various classification algorithms are applied on datasets using MATLAB tool and these datasets contain all relevant and irrelevant features. Table 5.4 shows a comparison of the accuracy, sensitivity, specificity, PPV and NPV of the diagnosis obtained from each individual model (LDA, DLDA, QDA, DQDA, SVM, and KNN), the KNN model appears to take the lead with accuracy 93.04%, followed by SVM of accuracy 68.41% as the runner-up model whereas DQDA with accuracy of 51.88% is the worst model in terms of overall performance for BUPA liver disorder dataset. Correspondingly, Table 5.5 shows the comparison of all algorithms for Liver damage dataset and here also KNN achieved high accuracy of 86.14% but DLDA achieved least accuracy of 28.05% and Table 5.6 represents the performance evaluation results of Hepatitis dataset where KNN model appears to take the lead with an accuracy of 96.77% and SVM achieved accuracy of 91.61% which is the second best accuracy. Thus, KNN is considered as the best algorithm for all the three mentioned datasets.

**Table 5.4** The comparison results of single models for BUPA liver disorder dataset.

| Single Model/Algorithm | | LDA | DLDA | QDA | DQDA | SVM | KNN |
|---|---|---|---|---|---|---|---|
| **Accuracy (%)** | **Training** | 65.58 | 58.7 | 59.42 | 51.45 | 69.93 | 100 |
| | **Testing** | 64.93 | 59.71 | 58.26 | 51.88 | 68.41 | **93.04** |
| **Sensitivity (%)** | **Training** | 69.83 | 73.28 | 88.79 | 83.62 | 70.69 | 100 |
| | **Testing** | 70.34 | 73.1 | 84.83 | 83.45 | 68.97 | 90.34 |
| **Specificity (%)** | **Training** | 62.5 | 48.13 | 38.12 | 28.13 | 69.37 | 100 |
| | **Testing** | 61 | 50 | 39 | 29 | 68 | 95 |
| **PPV (%)** | **Training** | 57.45 | 50.6 | 50.99 | 45.75 | 62.6 | 100 |
| | **Testing** | 56.67 | 51.46 | 50.2 | 46.01 | 60.98 | 92.91 |
| **NPV (%)** | **Training** | 74.07 | 71.3 | 82.43 | 70.31 | 76.55 | 100 |
| | **Testing** | 73.94 | 71.94 | 78 | 70.73 | 75.14 | 93.14 |

**Table 5.5** The comparison results of single models for Liver damage dataset.

| Single Model/Algorithm | | LDA | DLDA | QDA | DQDA | SVM | KNN |
|---|---|---|---|---|---|---|---|
| **Accuracy (%)** | **Training** | 28.04 | 27.42 | 32.58 | 31.13 | 53.25 | 100 |
| | **Testing** | 28.22 | 28.05 | 32.18 | 31.85 | 53.75 | **86.14** |
| **Sensitivity (%)** | **Training** | 36.36 | 37.7 | 64.75 | 69.67 | 52.66 | 100 |
| | **Testing** | 36.84 | 39.47 | 65.13 | 68.42 | 53.07 | 88.82 |
| **Specificity (%)** | **Training** | 67.03 | 71.35 | 49.31 | 44.63 | 55.64 | 100 |
| | **Testing** | 67.18 | 70.93 | 47.8 | 47.14 | 56.37 | 95.59 |
| **PPV (%)** | **Training** | 26.83 | 30.67 | 30.04 | 29.72 | 78.28 | 100 |
| | **Testing** | 27.32 | 31.25 | 29.46 | 30.23 | 78.95 | 87.1 |
| **NPV (%)** | **Training** | 76.01 | 77.31 | 80.63 | 81.41 | 29.15 | 100 |
| | **Testing** | 76.06 | 77.78 | 80.37 | 81.68 | 29.32 | 96.23 |

**Table 5.6** The comparison results of single models for Hepatitis dataset.

| Single Model/Algorithm | | LDA | DLDA | QDA | DQDA | SVM | KNN |
|---|---|---|---|---|---|---|---|
| **Accuracy (%)** | **Training** | 94.35 | 87.1 | N/A | N/A | 91.13 | 100 |
| | **Testing** | 90.32 | 85.16 | N/A | N/A | 91.61 | **96.77** |
| **Sensitivity (%)** | **Training** | 88.46 | 80 | N/A | N/A | 96.15 | 100 |
| | **Testing** | 84.38 | 81.25 | N/A | N/A | 96.88 | 93.75 |
| **Specificity (%)** | **Training** | 95.92 | 88.89 | N/A | N/A | 89.8 | 100 |
| | **Testing** | 91.87 | 86.18 | N/A | N/A | 90.24 | 97.56 |
| **PPV (%)** | **Training** | 85.19 | 64.52 | N/A | N/A | 71.43 | 100 |
| | **Testing** | 72.97 | 60.47 | N/A | N/A | 72.09 | 90.91 |
| **NPV (%)** | **Training** | 96.91 | 94.62 | N/A | N/A | 98.88 | 100 |
| | **Testing** | 95.76 | 94.64 | N/A | N/A | 99.11 | 98.36 |

**N/A**-There is no result available for QDA and DQDA. As the covariance matrix needs here to be a positive definite. There is an assumption that our data is generally represented by a multivariate probability distribution, which always has a positive definite covariance matrix unless one or more variables are exact linear combinations of the others. This problem can be resolved by performing dimensionality reduction approach on training data and then classify using few dimensions.

#### *5.3.2.2 Performance Evaluation of Classification models integrated with PCAs*

PCA is a feature extraction technique where on the basis of transformation function a set of new reduced features is created. In single models we have all the dimensions or features whether they are important or not but in order to keep only important features, we used PCA technique and this PCA has reduced the principal components by removing irrelevant feature combinations.

PCA is added to each of the single model forming integrated models PCA-LDA, PCA-DLDA, PCA-QDA, PCA-DQDA, PCA-SVM, and PCA-KNN to analyze the variation in results. Table 5.7 shows the results of PCA-added integrated models for BUPA liver disorder dataset in terms of accuracy, sensitivity, specificity, PPV and NPV where PCA-KNN achieved high accuracy of 91.88% and PCA-QDA achieved least accuracy of 53.91%. In Table 5.8 PCA-KNN has high prediction accuracy of 85.31% and PCA-LDA has least accuracy of 29.54% for liver damage dataset. In Hepatitis dataset also, PCA-KNN outruns all other PCA-added integrated models by achieving an accuracy of 98.71% but PCA-LDA and PCA-DLDA achieved least accuracy of 89.68% which has been shown in Table 5.9. PCA-KNN is considered as best integrated model for BUPA liver disorder dataset, Liver damage dataset and for Hepatitis dataset but worst performer algorithm is not same for all the datasets.

**Table 5.7** The comparison results of PCA-based integrated models for BUPA liver disorder dataset.

| **Integrated Model** | | **PCA-LDA** | **PCA-DLDA** | **PCA-QDA** | **PCA-DQDA** | **PCA-SVM** | **PCA-KNN** |
|---|---|---|---|---|---|---|---|
| **Accuracy (%)** | **Training** | 58.33 | 57.97 | 54.35 | 53.62 | 59.06 | 100 |
| | **Testing** | 59.42 | 59.13 | 53.91 | 54.49 | 58.26 | **91.88** |
| **Sensitivity (%)** | **Training** | 65.52 | 65.52 | 84.48 | 80.17 | 74.14 | 100 |
| | **Testing** | 67.59 | 67.59 | 82.07 | 80.69 | 75.17 | 89.66 |
| **Specificity (%)** | **Training** | 53.13 | 52.5 | 32.5 | 34.38 | 48.13 | 100 |
| | **Testing** | 53.5 | 53 | 33.5 | 35.5 | 46 | 93.5 |
| **PPV (%)** | **Training** | 50.33 | 50 | 47.57 | 46.97 | 50.89 | 100 |
| | **Testing** | 51.31 | 51.04 | 47.22 | 47.56 | 50.23 | 90.91 |
| **NPV (%)** | **Training** | 68 | 67.74 | 74.29 | 70.51 | 71.96 | 100 |
| | **Testing** | 69.48 | 69.28 | 72.04 | 71.72 | 71.88 | 92.57 |

**Table 5.8** The comparison results of PCA-based integrated models for Liver damage dataset.

| Integrated Model | | PCA-LDA | PCA-DLDA | PCA-QDA | PCA-DQDA | PCA-SVM | PCA-KNN |
|---|---|---|---|---|---|---|---|
| **Accuracy (%)** | **Training** | 29.28 | 31.13 | 31.13 | 32.58 | 51.29 | 100 |
| | **Testing** | 29.54 | 29.7 | 30.53 | 31.02 | 51.81 | **85.31** |
| **Sensitivity (%)** | **Training** | 48.76 | 50.82 | 67.21 | 73.77 | 50.17 | 100 |
| | **Testing** | 52.63 | 50 | 66.45 | 68.42 | 50.56 | 86.84 |
| **Specificity (%)** | **Training** | 58.52 | 60.88 | 47.66 | 42.7 | 55.28 | 100 |
| | **Testing** | 56.39 | 60.13 | 46.04 | 42.95 | 56.19 | 94.93 |
| **PPV (%)** | **Training** | 28.1 | 30.39 | 30.15 | 30.2 | 77.06 | 100 |
| | **Testing** | 28.78 | 29.57 | 29.19 | 28.65 | 77.62 | 85.16 |
| **NPV (%)** | **Training** | 77.45 | 78.65 | 81.22 | 82.89 | 28.02 | 100 |
| | **Testing** | 78.05 | 78.22 | 80.38 | 80.25 | 28.33 | 95.57 |

**Table 5.9** The comparison results of PCA-based integrated models for Hepatitis dataset.

| Integrated Model | | PCA-LDA | PCA-DLDA | PCA-QDA | PCA-DQDA | PCA-SVM | PCA-KNN |
|---|---|---|---|---|---|---|---|
| **Accuracy (%)** | **Training** | 89.52 | 91.94 | 96.77 | 91.94 | 91.13 | 100 |
| | **Testing** | 89.68 | 89.68 | 96.13 | 90.32 | 90.97 | **98.71** |
| **Sensitivity (%)** | **Training** | 80 | 92.31 | 100 | 96.15 | 92.31 | 100 |
| | **Testing** | 78.13 | 90.63 | 96.88 | 93.75 | 93.75 | 96.88 |
| **Specificity (%)** | **Training** | 91.92 | 91.84 | 95.92 | 90.82 | 90.82 | 100 |
| | **Testing** | 92.68 | 89.43 | 95.93 | 89.43 | 90.24 | 99.19 |
| **PPV (%)** | **Training** | 71.43 | 75 | 86.67 | 73.53 | 72.73 | 100 |
| | **Testing** | 73.53 | 69.05 | 86.11 | 69.77 | 71.43 | 96.88 |
| **NPV (%)** | **Training** | 94.79 | 97.83 | 100 | 98.89 | 97.8 | 100 |
| | **Testing** | 94.21 | 97.35 | 99.16 | 98.21 | 98.23 | 99.19 |

#### *5.3.2.3 Performance Comparison of Applied Techniques*

The motivation behind this work is the varied performances of machine learning algorithms for distinct datasets as observed from literature study. During literature review, it has been observed that a number of machine learning algorithms have been applied to various datasets. Few research claims that individual models are performing better. For example, E. M. Hashem et al. [11] analyze the performance of classification algorithms over two distinct datasets and finds that SVM is the best algorithm as compare to other algorithms. On the other side, few research claims that integrated models are performing better. For example, C. Chuang [13] presents BPN-CBR outperformed all single and other integrated models.Then for this study we took three different datasets and chooses a variety of machine learning algorithms and analyze the performances of all algorithms using different datasets in terms of accuracy, sensitivity, specificity, PPV and NPV.

The study examined several models of machine learning in search for an optimal method capable of performing more accurate and sensitive disease diagnosis. For example, in this study we have used different machine learning algorithms for three datasets to find the applications of each algorithm in terms of their performance parameters. Both individual and integrated models are developed to analyze their performances. In Table 5.4, Table 5.5 and Table 5.6 KNN model outperformed all other individual models for BUPA liver disorder dataset, Liver damage dataset and for Hepatitis dataset by achieving the highest accuracy. KNN is automatically non linear, linear and non-linear distribution data can easily be detected by KNN. It is completely non parametric approach where no assumptions can be made about the shape of the decision boundary. So when the decision boundary is extremely non-linear then KNN dominates other algorithms. When the data points are large like in all three datasets then also KNN outruns other algorithms. It is also sensitive to outliers and removing them before using KNN. It mostly works with numerical values. On the other side, SVM is the runner up model for all the three datasets and unable to achieve high accuracy because SVM needs to select good kernel function and appropriate hyper parameters that will allow for sufficient generalization performance but it is a difficult task. Moreover, SVM does not work well with large number of observations. In Table 5.4, DQDA is the worst performer as it has least accuracy as compare to other individual algorithms. DQDA accuracy is less because it does not perform well in case of large number of observations. In Table 5.5 and Table 5.6, DLDA has the least accuracy as decision boundary is not sufficiently flexible i.e. it is not a linear.

In Table 5.6, there is no result of QDA and DQDA, because the covariance matrix needs here to be a positive definite, there is an assumption that our data is generally represented by a multivariate probability distribution, which always has a positive definite covariance matrix unless one or more variables are exact linear combinations of the others. This problem can be resolved by performing PCA on training data and then classify using first few principal components. Thus, in Table 5.9 we have seen the results of QDA and DQDA after applying PCA.

In Table 5.7, Table 5.8 and Table 5.9 PCA-KNN has the highest accuracy as compare to other algorithms. After reducing dimensions, in Table 5.7 and Table 5.8 accuracy of PCA-KNN is not improving as compare to accuracy of individual KNN model in Table 5.4 and Table 5.5. Actually PCA does not have any correlation with

classification accuracy necessarily. Here 95% of the total variance corresponds to the first few principal components and the last principal component (contributes 5% variance) which we drop may have the actual ability to classify the data. Generally lower variance principal component associated with noise and there is an advantage to remove them but there is no guarantee of this. As a result of this accuracy of PCA-KNN is not improving here as compare to single KNN model. However, in Table 5.9 for Hepatitis dataset, accuracy of PCA-KNN is improving after reducing dimensions as compare to individual KNN model in Table 5.6. In this we have dropped those principal components which contain lot of noise and they do not have the ability to classify the data.

It is not necessary that these results will be same for all datasets. Some researchers claimed SVM as the best classification model but here SVM is not showing best accuracy. Also one researcher C. Chuang [13] presents integrated model outperformed all single and other integrated models but here in our study we have observed integrated model not necessarily performs better than all single and integrated models as shown in Table 5.7 and Table 5.8. Thus from this study we observe that performance of algorithm totally depends upon the dataset, their type of data, their number of observations, their dimensions and their decision boundary. From the above discussion we can also say that after dimensionality reduction it is not necessary that accuracy for that particular model will be improved, it may or may not be improved as it depends upon the principal components that we have removed or chopped down.

### 5.3.3 Experimental Results of Methodology II

In this study, classification is performed by case comparison. Leave-one-out is an evaluation method to measure the performance of an algorithm. It is a cross validation approach where training is performed on all data except for one point and prediction is made for that point. The matching cases to the new case can be found by using K-NN method. K-NN computation represents k number of matching cases from the prior case-base. For both the simple and integrated model, K-NN calculation is performed by changing the value of k from 1 to 9 by taking only odd numbers. After performing the experiment, the result predicts that 5-NN shows the best performance for the different cases. So we fix the value of k equal to 5 to find the 5 best matching cases to

the new case from the reference cases. First, a CBR model is applied to all the three datasets to evaluate the accuracies and then simultaneous weight optimization is done by GA to enhance the performance evaluation of an integrated model. The metric accuracy of a model is calculated using following equation:

$$\text{Accuracy} = \frac{\text{No. of correct predictions}}{\text{Total of all cases to be predicted}}$$

$$= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (15)$$

For liver disease prediction, the values of TP, TN, FP and FN are interpreted as follows:

TP= The CBR system decides the liver disease case, and domain expert decides a liver disease case.
TN= The CBR system decides not a liver disease case, and also the expert decides not a liver disease case.
FP= The CBR system decides a liver disease case, but the domain expert do not.
FN= the CBR system decides not a liver disease case, but the domain expert decides it is liver disease case.

#### *5.3.3.1 Performance Evaluation of CBR Model*

CBR model is applied and the accuracies are measured for the above discussed three liver disease datasets. After loading the entire dataset, five most matching cases are found for every case with the help of K-Nearest Neighbor. Table 5.10 contains the details of a number of cases lie between the different accuracies range. For example, in BUPA liver disorder dataset 2, 3, 45, 224, 68 and 3 cases lie between (>=0 and <5), (>=30 and <60), (>=60 and <65), (>=65 and <70), (>=70 and <75) and (>=75 and <85) accuracies range. In liver damage dataset 3, 106, 482, 9 and 6 cases lie between (>=0 and <5), (>=10 and <15), (>=15 and <20), (>=20 and <25) and (>=25 and <30) accuracies range. In hepatitis dataset 46, 75 and 4 cases lie between (>=85 and <90), (>=90 and <95), (>=95 and <100) accuracies range and 30 cases are having accuracy 100%. Table 5.11 shows an average accuracy which is 65.79, 17.32, and 87.09 for BUPA liver disorder dataset, Liver damage dataset, and Hepatitis dataset.

**Table 5.10** Case details of different datasets.

| Accuracies Range (%) | Number of BUPA liver disorder dataset cases | Number of liver damage dataset cases | Number of Hepatitis dataset cases |
|---|---|---|---|
| >=0 and <5 | 2 | 3 | -- |
| >=5 and <10 | -- | -- | -- |
| >=10 and <15 | -- | 106 | -- |
| >=15 and <20 | -- | 482 | -- |
| >=20 and <25 | -- | 9 | -- |
| >=25 and <30 | -- | 6 | -- |
| >=30 and <35 | 1 | -- | -- |
| >=35 and <40 | -- | -- | -- |
| >=40 and <45 | -- | -- | -- |
| >=45 and <50 | -- | -- | -- |
| >=50 and <55 | 1 | -- | -- |
| >=55 and <60 | 1 | -- | -- |
| >=60 and <65 | 45 | -- | -- |
| >=65 and <70 | 224 | -- | -- |
| >=70 and <75 | 68 | -- | -- |
| >=75 and <80 | 2 | -- | -- |
| >=80 and <85 | 1 | -- | -- |
| >=85 and <90 | -- | -- | 46 |
| >=90 and <95 | -- | -- | 75 |
| >=95 and <100 | -- | -- | 4 |
| equal to 100 | -- | -- | 30 |
| **Total Cases** | **345 cases** | **606 cases** | **155 cases** |

**Table 5.11** CBR Accuracy details of different datasets.

| Datasets | BUPA liver disorder dataset | Liver damage dataset | Hepatitis dataset |
|---|---|---|---|
| CBR Accuracy (%) | 65.79 | 17.32 | 87.09 |

**Table 5.12** GA-CBR Accuracy details of different datasets.

| Datasets | BUPA liver disorder dataset | Liver damage dataset | Hepatitis dataset |
|---|---|---|---|
| GA-CBR Accuracy (%) | 68.98 | 24.42 | 94.19 |

#### *5.3.3.2 Simultaneous weight optimization using GA (Integrated GA-CBR model)*

CBR model has been widely used in the medical field. In this study, it has been used to diagnose the presence of liver disease or not. However, as a contrast with other machine learning methods CBR model is scrutinized in view of low prediction performance. Retrieval of cases from prior case base should be effective in order to obtain better results. Optimization of weights of features and appropriate instance

selection simultaneously may lead to better performance than independent models. So here simultaneous optimization of weights and selection of appropriate instances for CBR is done by GA to improve the efficiency. The GA weight learning module is completely coordinated with the CBR.

The GA advances the weight structure. First, set the number of chromosomes to a small value i.e. 20 and set mutation rate to 0.05%. Every chromosome characterizes to the weights 1..N, as appropriate to a case base of cases each having N fields. The number of bits is 2 which represent 2*2=4 number of weight values. The weight 1 is represented by bit 00, weight 2 is represented by bit 01, weight 3 is represented by bit 10, and weight 4 is represented by bit 11. After setting all these parameters, run the CBR to evaluate each chromosome. The different genetic operators like crossover and mutation are applied to evaluate each new individual. Table 5.12 represents the accuracies of GA for different liver disease datasets which are giving better results as compare to CBR. The BUPA liver disorder dataset, liver damage dataset, and hepatitis dataset shows the accuracy of 68.98%, 24.42% and 94.19%. However, the accuracy result is best for hepatitis dataset i.e. 94.19%.

#### *5.3.3.3 Performance Comparison of Applied Techniques*

The motivation behind this work is the low performance shown by CBR technique. GA based CBR model has been developed to enhance its performance. GA is a heuristic solution search procedure motivated by natural evolution. It is a powerful and adaptable approach that can be connected to an extensive variety of learning and enhancement issues. In this optimization of weights of features and selection of suitable instances have been done simultaneously. This optimization is found to improve the retrieval results for complete CBR retrieval and its components. It is observed that optimization of these components at the same time leads to better results than separate optimization. As Table 5.11 shows the accuracy details of CBR whereas Table 5.12 shows the improved performance of CBR by an integrated GA-CBR model. This shows an effective impact of GA algorithm in the prediction results of CBR in liver disease. GA follows the concept of solution evolution by stochastically creating eras of solution population utilizing an objective function. GA is especially applicable to problems which are very large, discrete and non-linear in nature. They are applicable to features that add to the degree of complexity of

solution. They do not break easily in the presence of noise or even if the inputs changed slightly. In addition to, searching can takes place in large state-space, multi-model state-space and n-dimensional state-space. These are some applicable reasons which make this model to optimize the performance of CBR.

Moreover, this model shows large variations in accuracies in terms of datasets. Table 5.12 shows the highest accuracy of 94.19% for hepatitis dataset as compare to other datasets. Here, features of hepatitis dataset add to the degree of complexity of solution. In addition to this, hepatitis dataset is discrete and non-linear in nature. Due to these reasons GA-CBR model is showing highest accuracy for hepatitis dataset as compare to other datasets. It is observed that the prediction performance of CBR has been optimized by introducing our proposed GA-CBR model. The applications of GA are purely available in hepatitis dataset and for the reason that this dataset shows highest accuracy.

# CHAPTER 6
# CONCLUSION AND FUTURE SCOPE

Although there is a great advancement in medical field but diagnosis of liver disease is still an exigent task. As diagnosis of liver disease at an initial stage is very difficult because liver works properly even it is partially damaged. In this study to make the diagnosis more proficient and effectual, classification models developed to predict whether the liver disease is present or not.

In first methodology, KNN has best prediction accuracy in BUPA liver disorder dataset, Liver damage dataset and in Hepatitis dataset by comparing single method models. The PCA is integrated with single method models to improve the performance of classification models but PCA is not directly correlated with the classification accuracy. It is not necessary that after applying PCA accuracy will increase. Sometimes accuracy will drop because in PCA we sometimes drop that low variance principal component which is actually responsible for classification of data. By doing all literature review and by observing our results we can conclude that one particular algorithm can't show the high accuracy results for all liver related datasets i.e. there are different algorithms showing high prediction accuracy for different datasets. One more point need to be concluded here that integrated model not always performs better than individual model. Thus we can say that no algorithm is ideal. The performance of the algorithms totally depends upon the type of dataset, their number of observations, their dimensions and their decision boundary.

Some machine learning techniques are reprimanded in view of its confinement's like- poor clarification capacity of the outcomes and overfitting. CBR model has been applied to overcome all those limitations, but it has weak performance. CBR model is showing an accuracy of 65.79%, 17.32%, and 87.09% for BUPA liver disorder dataset, liver damage dataset and hepatitis dataset. We have proposed an integrated GA and CBR framework that upgrades weights of elements and select appropriate cases at the same time. This hybrid model decreases the noise and removes the imprecise cases which cause incorrect prediction of performance. Integrated GA-CBR model is showing an accuracy of 68.98%, 24.42% and 94.19% for three different datasets which are better as compare to single CBR model.

However, the hybrid GA-CBR model is not showing the better results as compare to the results of algorithms used in first methodology. As outcomes of first methodology shows that PCA-KNN has highest accuracy among all integrated algorithms for all datasets but integrated GA-CBR is not showing better accuracy as compare to them. We applied the CBR approach as it is data driven approach, very easy to apply, provides good explanation for the output and it removes the problem of overfitting which are the limitations of other applied algorithms. Our future work is centered on a few regions: the change in the representation of GA, the improvement on the distinctive phases of CBR.

# REFERENCES


[1] H. C. V. A. N. D. T. H. E. L. I. Ver, "An Overview of the Liver," no. April, 2015.

[2] P. Angulo, "Nonalcoholic Fatty Liver Disease," *N. Engl. J. Med.*, vol. 346, no. 16, pp. 1221–1231, 2002.

[3] D. Schuppan and N. H. Afdhal, "Liver cirrhosis," *The Lancet*, vol. 371, no. 9615. pp. 838–851, 2008.

[4] A. T. Duddempudi and D. E. Bernstein, "Hepatitis B and C," *Clinics in Geriatric Medicine*, vol. 30, no. 1. pp. 149–167, 2014.

[5] J. L. Newton and D. E. J. Jones, "Managing systemic symptoms in chronic liver disease," *J. Hepatol.*, vol. 56, no. SUPPL. 1, 2012.

[6] A. Gulia, R. Vohra, and P. Rani, "Liver Patient Classification Using Intelligent Techniques," vol. 5, no. 4, pp. 5110–5115, 2014.

[7] A.Branch, and I.Azad, "Using algorithms to predict liver disease Classification," *Electronics Information & Planning*, vol. 3, pp. 255-259, 2015.

[8] B. V. Ramana, P. M. Surendra, P. Babu, and P. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *International Journal of Database Management Systems*, vol. 3, no. 2, pp. 101–114, 2011.

[9] A. S. Aneeshkumar, "Estimating the Surveillance of Liver Disorder using Classification Algorithms," *International Journal of Computer Applications* , vol. 57, no. 6, pp. 39–42, 2012.

[10] S. Karthik, A. Priyadarishini, J. Anuradha, and B. K. Tripathy, "Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types," *Advances in Applied Science Research*, vol. 2, no. 3, pp. 334–345, 2011.

[11] E. M. Hashem and M. S. Mabrouk, "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis," *American Journal of Intelligent Systems*, vol. 4, no. 1, pp. 9–14, 2014.

[12] R. Lin and C. Chuang, "A hybrid diagnosis model for determining the types of the liver disease," *Comput. Biol. Med.*, vol. 40, no. 7, pp. 665–670, 2010.

[13] C. Chuang, "Artificial Intelligence in Medicine Case-based reasoning support for liver disease diagnosis," *Artif. Intell. Med.*, vol. 53, no. 1, pp. 15–23, 2011.

[14] P. Tamije Selvy , V. Palanisamy and S. Elakkiya, " Evaluation of Classification

Algorithms for Disease Diagnosis," *Journal of Global Research in Computer Science*, vol. 4, no. 4, pp. 77–81, 2013.

[15] A. Singh and B. Pandey, "Diagnosis of Liver Disease by Using least Squares Support Vector Machine Approach," *Int. J. Healthc. Inf. Syst. Informatics*, vol. 11, no. 2, pp. 62–75, Apr. 2016.

[16] X. X. Geng, R. G. Huang, J. M. Lin, N. Jiang, and X. X. Yang, "Transient Elastography in Clinical Detection of Liver Cirrhosis : A Systematic Review and Meta - analysis," vol. 22, no. 4, pp. 294–303, 2016.

[17] T. Orczyk and P. Porwik, "Liver Fibrosis Diagnosis Support System Using Machine Learning Methods BT - Advanced Computing and Systems for Security: Volume 1," R. Chaki, A. Cortesi, K. Saeed, and N. Chaki, Eds. New Delhi: Springer India, 2016, pp. 111–121.

[18] L. Ozyilmaz and T. Yildirim, "Artificial Neural Networks for Diagnosis of Hepatitis Disease," pp. 586–589.

[19] I. Campus, "Diagnosis Of Liver Disease Induced By Hepatitis Virus Using Artificial Neural Networks," pp. 8–12, 2011.

[20] G. S. Uttreshwar, "Hepatitis B Diagnosis Using Logical Inference And Generalized Regression Neural Networks," *Advance Computing Conference*, no. March, pp. 6–7, 2009.

[21] R. Stoean, C. Stoean, M. Lupsor, H. Stefanescu, and R. Badea, "Artificial Intelligence in Medicine Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C," *Artif. Intell. Med.*, vol. 51, no. 1, pp. 53–65, 2011.

[22] A. C. Approach, "Prediction of the Degree of Liver Fibrosis Using Different Pattern Recognition Techniques," *Biomedical Engineering Conference*, pp. 1–5, 2010.

[23] A. G. Floares, "Chronic Hepatitis C and B based on i-BiopsyTM," pp. 855–860, 2009.

[24] D. Lu, Y. Wu, G. Harris, and W. Cai, "Iterative mesh transformation for 3D segmentation of livers with cancers in CT images," *Comput. Med. Imaging Graph.*, vol. 43, pp. 1–14, Jul. 2015.

[25] C. T. C. Arsene and P. J. Lisboa, "Bayesian Neural Network Applied in Medical Survival Analysis of Primary Biliary Cirrhosis," Computer Modelling and Simulation, pp. 81–85, 2012.

[26] X. Zhang, X. Gao, B. J. Liu, K. Ma, W. Yan, L. Liling, H. Yuhong, and H. Fujita, "Effective staging of fibrosis by the selected texture features of liver: Which one is better, CT or MR imaging?," *Comput. Med. Imaging Graph.*, vol. 46, Part 2, pp. 227–236, Dec. 2015.

[27] D. Li, C. Liu, and S. C. Hu, "Artificial Intelligence in Medicine A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," *Artif. Intell. Med.*, vol. 52, no. 1, pp. 45–52, 2011.

[28] B. Pandey and A. Singh, "Intelligent techniques and applications in liver disorders : A survey," *Int. J. of Biomedical Engineering and Technology*, vol. 16, no. 1, pp. 27-70, 2014.

[29] S. Petrovic, N. Mishra, and S. Sundar, "A novel case based reasoning approach to radiotherapy planning," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 10759–10769, 2011.

[30] V. E. Ekong, U. G. Inyang, and E. A. Onibere, "Intelligent Decision Support for Depression diagnosis based on Neuro-fuzzy CBR Hybrid Intelligent Decision Support System for Depression Diagnosis Based on Neuro-fuzzy-CBR Hybrid," *Modern Applied Science*, vol. 6, no. 7, pp. 79- 88, 2015.

[31] D. A. Sharaf-el-deen and F. Ibrahim, "A New Hybrid Case-Based Reasoning Approach for Medical Diagnosis Systems," *Journal of Medical Systems*, vol. 38, no. 9, pp. 1-11, 2014.

[32] Z. Yin, Z. Dong, X. Lu, S. Yu, X. Chen, and H. Duan, "A clinical decision support system for the diagnosis of probable migraine and probable tension-type headache based on case-based reasoning," *The Journal of Headache and Pain*, vol. 16, no. 29, pp. 1-9, 2015.

[33] R. M. Saraiva, J. Bezerra, M. Perkusich, H. Almeida, and C. Siebra, "A Hybrid Approach Using Case-Based Reasoning and Rule-Based Reasoning to Support Cancer Diagnosis : A Pilot Study," *Studies in health technology and informatics*, vol. 216, pp. 862–866, 2015.

[34] A. Ghaheri, S. Shoar, M. Naderan, and S. S. Hoseini, "The Applications of Genetic Algorithms in Medicine," *Oman Medical Journal*, vol. 30, no. 6, pp. 406–416, 2015.

[35] C. Wu, W. Lee, Y. Chen, C. Lai, and K. Hsieh, "Expert Systems with Applications Ultrasonic liver tissue characterization by feature fusion," *Expert System Applications*, vol. 39, no. 10, pp. 9389–9397, 2012.

[36] P. Pal, S. Tomar, and R. Singh, "Evolutionary Continuous Genetic Algorithm for Clinical Decision Support System," *African Journal of Computing & ICT*, vol. 6, no. 1, pp. 127–140, 2013.

[37] E. Sreedevi and P. M. Padmavathamma, "A Threshold Genetic Algorithm for Diagnosis of Diabetes using Minkowski Distance Method," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 4, no. 7, pp. 5596–5601, 2015.

[38] D. A. Antony and G. Singh, "Dimensionality Reduction using Genetic

Algorithm for Improving Accuracy in Medical Diagnosis," *I.J. Intelligent Systems and Applications*, vol. 1, pp. 67–73, 2016.

[39] H. Kahramanli and N. Allahverdi, "Extracting rules for classification problems: AIS based approach," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10494–10502, 2009.

[40] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, 2007.

[41] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4. pp. 433–459, 2010.

# APPENDIX

## ABBREVIATIONS

1. NAFLD: Non-Alcoholic Fatty Liver Disease
2. LDA: Linear Discriminant Analysis
3. DLDA: Diagonal Linear Discriminant Analysis
4. QDA: Quadratic Discriminant Analysis
5. DQDA: Diagonal Quadratic Discriminant Analysis
6. SVM: Support Vector Machine
7. KNN: K-Nearest Neighbor
8. PCA: Principal Component Analysis
9. CBR: Case Based Reasoning
10. GA: Genetic Algorithm
11. UCI: University of California, Irvine
12. ILPD: Indian Liver Patient Dataset
13. NBC: Naïve Bayes Classification
14. ANN: Artificial Neural Network
15. LEM: Learn By Example
16. MLP: Multilayer Perceptron
17. CART: Classification and Regression Tree
18. BPN: Back Propagation Neural Network
19. DA: discriminant Analysis
20. LR: Logistic Regression
21. RBF: Radial Basis Function
22. SOM: Self Organizing Map
23. PPV: Positive Prediction Value
24. NPV: Negative Prediction Value
25. TP: True Positive
26. TN: True Negative
27. FP: False Positive
28. FN: False Negative

# PUBLICATONS

**Paper Accepted in Int. J. of E-Health and Medical Communications (IJEHMC) Journal**

Sakshi Takkar, Aman Singh and Babita Pandey, "Application of Machine Learning Algorithms to a Well Defined Clinical Problem: Liver Disease", *Int. J. of E-Health and Medical Communications*, vol. 8, no. 4, 2017.

**Paper Submitted to Int. J. of Performability Engineering (IJPE) Journal – Under Review**

Sakshi Takkar and Aman Singh, "Impact of Genetic Optimization on the Prediction Performance of Case Based Reasoning Algorithm in Liver Disease", *Int. J. of Performability Engineering*.