INTELLIGENT TECHNIQUES FOR THE DIAGNOSIS OF LIVER DISEASE

Thesis Submitted to

LOVELY PROFESSIONAL UNIVERSITY

For the award of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By

Aman Singh
Registration Number: 41200414

Supervised By

Dr. Babita Pandey nee Shukla
Associate Professor

FACULTY OF TECHNOLOGY AND SCIENCES

LOVELY PROFESSIONAL UNIVERSITY

PUNJAB

2017

DECLARATION

I hereby declare that the thesis entitled "Intelligent Techniques for the Diagnosis of Liver Disease" submitted by me for the Degree of Doctor of Philosophy in Computer Science and Engineering is the result of my original and independent research work carried out under the guidance of Dr. Babita Pandey, Associate Professor, Department of Computer Applications, Lovely Professional University, Punjab, and it has not been submitted for the award of any degree, diploma, associateship, fellowship of any University or Institution.

Aman Singh
(Registration Number: 41200414)
School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab-144411,
India
Date: June 27, 2017

CERTIFICATE

This is to certify that the thesis entitled "Intelligent Techniques for the Diagnosis of Liver Disease" submitted by Aman Singh for the award of the degree of Doctor of Philosophy in Computer Science and Engineering, Lovely Professional University, is entirely based on the work carried out by him under my supervision and guidance. The work reported, embodies the original work of the candidate and has not been submitted to any other university or institution for the award of any degree or diploma, according to the best of my knowledge.

Dr. Babita Pandey
Associate Professor
Department of Computer Applications,
Lovely Professional University,
Phagwara, Punjab-144411,
India
Date: June 27, 2017

ACKNOWLEDGEMENT

No creation in this world is a solo effort. Like every fine work, even this thesis is also a collaborative endeavor. First and foremost, I would like to thank the almighty for giving me enough mental strength and belief in completing this research work successfully. I wish to express my deep sense of gratitude and thanks to my supervisor Dr. Babita Pandey, for her continued encouragement and invaluable suggestions during this work. She has been actively interested in my work and has always been available to advise me. I am very grateful for her patience, motivation, enthusiasm and immense knowledge.

I wish to thank my Ph.D. doctoral committee members for providing valuable comments and suggestions during the course of this research work. I wish to express my sincere thanks to the management and staff of Division of Research and Development, Lovely Professional University for extending the support and facilities to carry out my research work. I want to express my gratitude towards my parents for their constant care and support as they always encouraged me to never ever give up**.** I am indebted to my friends whose works have always been a great source of inspiration for me. Special thanks to Divya Anand for her constant encouragement and help throughout this work. Lastly, I offer my regards to all of those who supported me in any respect during the completion of the research.

I thank and dedicate this effort to ALMIGHTY.

ABSTRACT

Intelligent techniques based diagnostic systems have played a vital role in medicine. From statistical techniques to data mining algorithms to neural networks, all these have been widely deployed on medical records for predicting the sickness. Due to increasing vagueness and complexities in health examination data, deriving intelligible information becomes a major challenge for physicians. This challenge could lead to imprecise assessment of the disease and would further direct inaccurate treatment to patients. Therefore to avoid these uncertainties in interpretation of multifaceted data up to a feasible extent, medical professionals employ intelligent techniques based prediction models. Like for other health complications, intelligent techniques have also shown significant performance in the diagnosis and classification of liver disease.

Liver is the largest internal organ in a human body which performs numerous metabolic functions. It filters blood, aids in digestion of fats, makes proteins for blood clotting and most importantly detoxifies harmful chemicals. Liver has a vital importance to life but improper functioning of it may cause serious health consequences. Liver disease is usually caused by inherited disorders, contaminated food, damaged hepatocytes infected with viruses, bacteria or fungi, excessive fat accumulation, and excessive consumption of alcohol or drugs. It is a serious area of concern in the universal set of medicine and is becoming the leading cause of death in India, as well as in other countries around the globe. Ability of liver to resist early detection, as it functions normally even when partially damaged, makes the disease even more alarming because by then it might have suffered eternal damage. This indicates that an early diagnosis of liver disease is a necessity so that in time treatment can be initiated. During diagnosis, analyzing complex medical records of patients may lead to erroneous evaluation and may stretch the decision time of doctors. To overcome these obstacles, computational models are developed using a variety of intelligent techniques which eventually assists the physicians in the diagnostic process.

The aim of this thesis is to build intelligent techniques based computational models for the diagnosis and classification of liver disease, and to analyze their performance using statistical parameters. The motivation behind this work is to assist physicians in liver disease evaluation process, to overcome liver biopsy up to a possible extent, to efficiently analyze complex and

ambiguous health examination data of patients, and to reduce the cost, time and effort needed. The intelligent models are developed for identifying liver disease, predicting degree of liver damage, classifying primary biliary cirrhosis, diagnosing hepatitis disease, and classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. These models are built using various categories of intelligent algorithms include dimensionality reduction methods, clustering techniques, instance based methods, decision trees, rule system methods, and ensemble learning approaches. Experimental results prove the credibility of proposed computational models in performing non-invasive automatic diagnosis and classification of liver disease. It is observed that the models have the capability of assisting physicians in examining patient liver health by making most efficient use of limited resources. For future perspective, these intelligent systems can also be deployed for predicting human disease other than liver.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2-LNN	Two-Level Neural Network
AIS	Artificial Immune System
AIRS	Artificial Immune Recognition System
AHP	Analytic Hierarchy Process
AMBC	Associative Memory Based Classifier
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ALB	Albumin
ALT	Alanine Transaminase
ALK	Alkaline Phosphatase
ALP	Alkaline Phosphatase
ARD	Automatic Relevance Determination
ART	Adaptive Resonance Theory
AST	Aspartate Transaminase
AWAIS	Attribute Weighted Artificial Immune System
BPN	Back-propagation Neural Network
BUN	Blood Urea Nitrogen
BNNF	Bayesian Network with Naive Dependence and Feature Selection
CART	Classification and Regression Tree
CBR	Case Based Reasoning
CBP	Complex Back-Propagation
CCEA	Cooperative Coevolutionary Algorithm
CRTNN	Creatinine
CT	Computed Tomography
CMAC	Cerebellar Model Articulation Controller
COG	Centre of Gravity
CSFNN	Conic Section Function Neural Network
CSCBR	Cost-Sensitive Case Based Reasoning
CVNN	Complex-Valued Artificial Neural Network
DA	Discriminatory Analysis
DB	Direct Bilirubin

DLDA	Diagonal Linear Discriminant Analysis
DQDA	Diagonal Quadratic Discriminant Analysis
DIMLP	Discretized Interpretable Multi-layer Perceptron
DM	Data Mining
DEA	Differential Evolution Algorithm
EDC	Evolutionary Data Classification
EL	Ensemble Learning
ESVM	Evolutionary-powered Support Vector Machine
FL	Fuzzy Logic
FCM	Fuzzy C-Means Clustering
FDA	Fisher Discriminant Analysis
FDT	Fuzzy Decision Tree
FIS	Fuzzy Inference System
FLR	Fuzzy Lattice Reasoning
FMLP	Fuzzy Multilayer Perceptron
FNN	Fuzzy Neural Network
FFNN	Feed-Forward Neural Network
FLNN	Functional Link Neural Network
FFT	Fast Fourier Transform
GA	Genetic Algorithm
GPT	Glutamic Pyruvic Transaminase
GGT	Gamma-Glutamyl Transferase
GGP	Generic Genetic Programming
GNNF	Generalized Regression Neural Network
GOT	Glutamic Oxalacetic Transaminate
GRNN	Generalized Regression Neural Network
HAV	Hepatitis A virus
HBV	Hepatitis B virus
HC	hierarchical Clustering
HCV	Hepatitis C virus
HDV	Hepatitis D virus
HEV	Hepatitis E virus
HGV	Hepatitis G virus
HDL	High Density Lipoprotein

IT	Intelligent Techniques
KLD	Kullback-Leibler Divergence
KM	K-means clustering
KNN	K Nearest Neighbor
KPCA	Kernel Principal Component Analysis
LR	Logistic Regression
LS	Least Squares
LDA	Linear Discriminant Analysis
LDH	Lactate Dehydrase
LM	Levenberg-Marquardt
LMS	Least Mean Square
LSSVM	Least Squares Support Vector Machine
LVQ	Learning Vector Quantization
MAE	Mean Absolute Error
MCV	Mean Corpuscular Volume
MCH	Mean Corpuscular Haemoglobin
McRBFN	Meta-cognitive Radial Basis Function Network
MDL	Minimal Description Length
MDM	Minimum Deviation Method
MLP	Multilayer Perceptron
MLPBPNN	Multilayer Perceptron Backpropagation Neural Network
MOEA	Multi-objective Evolutionary Learning Algorithm
MLRA	Multivariate Logistic Regression Analysis
MRI	Magnetic Resonance Imaging
MTD	Megatrend Diffusion
MSE	Mean Square Error
NB	Naive Bayes
NMIFS	Normalized Mutual Information Feature Selection
NPV	Negative Predictive Value
PBC	Primary Biliary Cirrhosis
PBL	Projection Based Learning
PCA	Principal Component Analysis
PLANN	Partial Logistic Artificial Neural Network
PNN	Probabilistic Neural Network

PSD	Power Spectral Densities
PSO	Particle Swarm Optimization
PPV	Positive Predictive Value
QP	Quadratic programming
QDA	Quadratic Discriminant Analysis
ROI	Region of Interest
RMSE	Root Mean Squared Error
RP	Resilient Propagation
RBF	Radial Basis Function
RBR	Rule Based Reasoning
RDP	Recursive Deterministic Perceptron
RF	Random Forest
SA	Simulated Annealing
SBMAC	Sub-population Based Max-Mean Arithmetical Crossover
SCG	Scaled Conjugate Gradient
STFT	Short Time Fourier Transform
SOM	Self Organizing Map
SMO	Sequential Minimal Optimization
SRA	Stepwise Regression Analysis
SSE	Sum Squared Error
SVM	Support Vector Machine
TACO	Touring Ant Colony Optimization
TB	Total Bilirubin
TP	Total Protein
TPR	True Positive Rate
TVMOPSO	Time Variant Multi-objective Particle Swarm Optimization
TVM	Time-Variant Mutation
VFI	Voting Feature Interval
VIF	Variance Inflationary Factor

Chapter 1

Introduction

This chapter provides an overview of the research presented in the thesis. Section 1.1 discusses the medical background. Section 1.2 elaborates the significance of intelligent techniques. Section 1.3 details the motivation behind the project, and Section 1.4 gives the outline of thesis containing brief introduction to the content of every chapter. The detailed description is as follows.

1.1 Medical Background

A healthy liver leads to healthy life. Liver is the largest internal organ in a human body weighing about 3 pounds [1]. It is reddish in color and has a notable property of self regeneration to its original size and shape. It performs various metabolic functions like filtering blood, producing bile, assisting in fat digestion, making proteins for blood clotting, metabolizing drugs, storing glucose, and most importantly detoxifying harmful chemicals [2]. Its neighboring organs are gallbladder, stomach, colon, pancreas and kidney. Liver has a vital importance to life but improper functioning of it may cause serious health consequences. Liver disease is generally caused by inherited disorders, contaminated food, damaged hepatocytes infected with viruses, bacteria or fungi, excessive fat accumulation, and excessive consumption of alcohol or drugs [3]. Early symptoms of the disease are loss of appetite, diarrhoea, nausea, fatigue or weight loss. As time passes, symptoms become more solemn like jaundice, hair loss, swollen abdomen, oedema, muscle wasting, memory problems, bleeding and bruising. Liver disease can be acute (for short time) or chronic (for long time) that can put human life at risk [4]. Severity of the disease may begin from a healthy individual to viral hepatitis infection, to cirrhosis and more seriously to liver cancer. Liver disease are categorized in more than hundred types out which a few common include viral hepatitis, autoimmune hepatitis, neonatal hepatitis, fibrosis, cirrhosis, primary hepatoma, liver tumor, liver abscess, cholelithiasis, primary biliary cirrhosis, wilson disease, alcoholic liver damage and non-alcoholic fatty liver damage.

Talk about organ failure and people immediately recall kidney disease. On contrary, there is no such alertness about liver disease and its failure despite the fact that it is one of the leading cause of mortality around the globe. It is a serious area of concern in the universal set

of medicine. Presence of liver disease is steadily increasing over the years irrespective of age, sex, region or race. It has been ranked as the fifth most common cause of death by national statistics. It is persistently listed as one of the top ten fatal diseases worldwide costing millions of lives every year [2]. Approximately 50% of the people are affected by liver diseases. Statistics estimate that 30 million people in United States have liver disease or we can say that 1 in 10 Americans, more than 3 million people have some form of liver disease or 1 in 11 Canadians, approximately 15,000 children are hospitalized every year with liver disorder in Canada, 29 million people in the European union still suffer from a chronic liver condition [5]–[7]. Cirrhosis presence is between 4.5% and 9.5% of general population and is estimated to increase rapidly which would make it the 12^{th} leading cause of death in 2020. According to WHO mortality database 1,70,000 deaths per year occur due to cirrhosis and 47,000 deaths due to liver cancer in Europe. In India, approximately 2 lakh people die of incurable liver disease of which 25,000 can be saved by transplants every year. European countries used to have high mortality rates with alcohol as a main cause. Today even Asian countries like India do have chronic liver disease because of excessive alcohol consumption. In United States, more than 30 million people have some form of liver disease and it is the second leading cause of mortality amongst all digestive disease. Presence of liver cancer in patients with cryptogenic cirrhosis has ranged from 6.9-29%. It is the 5^{th} and 7^{th} most common cancer in men and women respectively. Approximately 1 in 30,000 individual has wilson disease around the world with higher incidence in Asian countries. Fatty liver presence is estimated to be 9-32% in Indian population [8], one third of adults in the United States [9], 20-30% in western countries [10] and 10-24% worldwide [11] with vast majority undiagnosed and is still increasing year-on-year. Globally, alcoholic liver disease represents 9.5% of alcohol related disorder, 50 million people would be effected with chronic liver disease, hepatitis C infection is presented in 130-150 million people, more than 240 million people have chronic (long-term) liver infections, about 6,00,000 and 7,45,000 people die every year due to the consequences of hepatitis B and liver cancer respectively, 3,50,000 to 5,00,000 people die each year from hepatitis C-related liver diseases [12], [13]. All these facts are underestimated as almost a third of people remain asymptomatic.

Liver resists early detection, as it functions normally even when partially damaged, makes the disease even more alarming because by then it might have suffered eternal damage. This indicates that an early diagnosis of liver disease becomes a necessity so that in time treatment can be initiated [14]. Various modes of liver diagnosis are liver biopsy, image scan

(ultrasonography, computed tomography, magnetic resonance imaging etc.), liver function tests, medical history and physical examination [15]–[17]. Liver biopsy is still the gold standard method used to detect and characterize liver disease but has important sample error issues and subjectivity in the interpretation. Furthermore, it is an invasive method which can also raise a risk of complications if the mode of sampling is not appropriate. Physical examination and medical history do not replace other diagnostic procedures as physical indications are normal unless the damage is severe. It rather complements the diagnostic decision. Liver function tests majorly help in examining liver injury. Key parameters in the test include albumin, alkaline phosphatase, total proteins, alanine aminotransferase, aspartate aminotransferase, bilirubin, gamma-glutamyl transferase, prothrombin time, triglycerides platelet count and so on. These parameters indicate the liver damage but not specifically the type of damage. Few more biomarkers tests and physical symptoms are needed to be known to confirm type of liver disease. Magnetic resonance imaging scan has also become a popular mode of evaluation for liver classification but it involves a long procedure time and high cost. Ultrasonography is unable to differentiate between benign and malignant liver lesions and it cannot detect small hepatic lesions as it fails in penetrating air or bone. Computer tomography uses iodinated contrast material which is restricted in patients with renal insufficiency. These scans are limited to axial planes. Imaging techniques do have presence of structure noise in images which makes it difficult for medical expert to interpret precisely. Imaging devices generate huge amount of data that increases the chance of error occurrence. Clinical interpretations from a collection of symptoms, risk factors, laboratory examination tests and other vital examination figures is a highly demanding task in liver diagnosis. The task becomes even more complex if the existing figures are fuzzy. It also stretches the decision time of clinicians even if they are experienced, and if they are novice then it may take years for them to gain substantial expertise in analyzing these uncertain medical records of patients. Moreover, the accurate diagnosis is still not guaranteed as humans are prone to errors no matter whatever may the reason be like abundant clinical workload or a poor health. With the involvement of multiple evaluation modes, the definitive diagnostic method is difficult to decide as each method has their own merits and demerits. The search for an ideal non-invasive method for liver disease prediction has not been accomplished yet and experts still rely broadly on image scans and liver biopsy. Hence, to interpret multifaceted datasets, to avoid clinical inexperience, to reduce high cost and to minimize the evaluation time, this study presents computer-aided diagnostic systems build using a diversity of intelligent techniques for non-invasive diagnosis and classification of liver disease. Five distinct liver

health examination datasets of different dimensionality and structures are taken for experimentation. Many a times, the medical systems have shown the capability of replacing biopsy and imaging methods which furthermore reduces the diagnostic cost for patients.

1.2 Significance of Intelligent Techniques

"Intelligent techniques (ITs)" is a term refers to methods or algorithms which can be applied to almost any data problem. ITs enable computer systems to think and understand, to gather and incorporate domain knowledge, to learn from acquired knowledge, to apply knowledge and experience for manipulating the environment, to conclude situation with fuzziness and uncertainty, to recognize and infer in rational ways, to retort swiftly and effectively to new situations, to distinguish the relative importance of different elements in a state, to alter their conduct and respond to changes in the outer environment and to make sense out of vague or incongruous information. ITs generate solutions to large and complex problems; by capturing knowledge, discovering hidden patterns and relations in high dimensions data; that are too hard for humans to understand. ITs provide potential solution in domains like patter recognition, natural language processing, data mining, expert systems and image recognition. Its major application areas are medical, human resource management, planning, business, manufacturing and web services. ITs solve problems which numerical means unable to solve alone. These techniques exist because the problems are more complex and ambiguous in real time. ITs deal with huge dimensions of data for developing finely tuned predictor functions. These techniques can model a problem in different ways based on the interaction with input data. An intelligent technique has the ability to adopt different learning styles in order to get finest results. These styles include supervised learning, unsupervised learning and reinforcement learning [18], [19]. The first aforesaid learning has an input data and a target variable in which the training process continues until the model achieves maximum possible accuracy. A few examples of supervised learning include logistic regression, k-nearest neighbor, decision tree and random forest algorithm. The second learning style has unlabelled input data and doesn't have a target variable to predict. It generally works to cluster data into different groups for specific purpose. A few examples include k-means clustering, hierarchical clustering and apriori algorithm. Its model is prepared by deducing structures. In third style, the input data consists of labelled and unlabelled instances and the system trains itself using trial and error to make accurate decisions. A common example of reinforcement is markov decision process. Intelligent

techniques can be grouped into several categories include regression algorithms, regularization algorithms, instance-based algorithms, bayesian algorithms, association rule learning algorithms, artificial neural network algorithms, clustering algorithms, deep learning algorithms, decision tree algorithms, dimensionality reduction algorithms and ensemble learning algorithms. Regression algorithms use measure of error for refinement in the predictions. Instance based algorithms build the model of training data and then do prediction for new data using similarity measures between instances. Regularization algorithms are the extension to regression method that works on complexity measure, less complexity the model have, the more preferred it is. Decision tree algorithms represent information in tree structure based on actual values of features in data. Bayesian algorithms build the model for classification using bayes theorem. Clustering algorithms organize the instances into groups using centroid-based and hierarchal approaches. Association rule learning algorithms present relationships between attributes, in large multidimensional datasets, in form of rules. Artificial neural networks work by selecting data, creating and training a network, validating and testing the targets and evaluating the performance using confusion matrices and mean square error. Deep learning methods are the extensions to artificial neural networks which build larger and complex neural network structures. Dimensionality reduction methods perform feature selection and feature extraction to describe data using less information. Ensemble learning algorithms are the combination of two or more independently trained techniques whose predictions are combined to make a final prediction. The most popular regression algorithms are stepwise regression, linear regression, locally estimated scatterplot smoothing, ordinary least squares regression, logistic regression and multivariate adaptive regression splines; instance based algorithms are k-nearest neighbor, locally weighted learning, self organizing map and learning vector quantization; regularization algorithms are elastic net, ridge regression, least-angle regression and least absolute shrinkage and selection operator; decision tree algorithms are conditional decision trees, iterative dichotomiser 3, classification and regression tree, C4.5 and C5.0, decision stump, M5 and chi-squared automatic interaction detection; bayesian algorithms are bayesian network, naive bayes, multinomial naive bayes, gaussian naive bayes, bayesian belief network and averaged one-dependence estimators; clustering algorithms are expectation maximization, k-means, k-medians and hierarchical clustering; association rule learning algorithms are apriori and eclat algorithms; artificial neural network algorithms are hopfield, perceptron, back-propagation and radial basis function networks; deep learning algorithms are stacked auto-encoders, deep belief networks, deep boltzmann machine and convolutional neural network; dimensionality reduction

algorithms are linear discriminant analysis, principal component analysis, partial least squares regression, flexible discriminant analysis, mixture discriminant analysis, sammon mapping, projection pursuit, principal component regression, multidimensional scaling and quadratic discriminant analysis; and ensemble learning algorithms are random decision forest, adaptive boosting, bootstrapped aggregation, gradient boosted regression trees, boosting, gradient boosting machines and stacked generalization.

Computer-aided diagnostic systems build using intelligent techniques give finest decisions in shortest time possible and are considered more accurate as these techniques are data driven. These systems screen and filter the overflow of information, data and knowledge which eventually provide effective results. Human decisions are very subjective as it depends on individual judgment and preference. Human mind do have limitations of recalling crucial details of the problem. Intelligent techniques based systems produce fair and consistent decisions based on the learning and reasoning capabilities. It extends the support to explore information, to organize compound objects, to recognize the meaning of information available, and to improve human decisions rather than replacing the judgments. These techniques provide computers the ability to learn from N number of experiences with respect to some task T and to act suitably in an indecisive environment for increasing the likelihood of success. It makes the systems self modifying and highly automated which continue to improve over time as it learns with more data. It reduces the need of human intervention. Cross validation methods like hold-out, leave-one-out, resubstitution and k-fold cross validation are used to authenticate the prediction performances of these systems [20],[21]. Performance is measured using various statistical parameters like accuracy, sensitivity, specificity, positive predicting value and negative predicting values. The key constraints of ITs are: impossibility of getting perfect accuracy and requirement of labeled data in huge amount for better training. In general, no intelligent algorithm dominates all others on all given problems. Selection of an IT depends on the size, quality and structure of data. It is not viable to decide which IT will do finest predictions before trying them. Sometimes a specific technique naturally fits the problem but the best way is to try numerous techniques for finding the best fit.

1.3 Motivation

Modern era of computing has stretched its reach to the intensive and efficient usage of intelligent techniques in medicine. Owing to the uncertainty and complexity in medical

datasets, deriving comprehensible information becomes a major challenge for physicians. This challenge can lead to erroneous diagnosis of the disease, which would further lead to improper treatment. Therefore, it would be favorable for patients if medical experts cross check their assessment with the help of computer-aided diagnostic systems. These systems are developed using intelligent techniques which resourcefully scrutinize complex and ambiguous medical data. Implementing ITs for predicting liver disease is acting as a catalyst in overcoming the overheads and problems faced by clinicians [22]. ITs effectively prevail over the inadequacies, help in obtaining better classifying performances and make the systems adaptable. ITs decrease the probability of occurrence of diagnostic errors, and reduces the cost, time and effort needed. Now a days, these systems are being developed using integrated methodologies in which two or more intelligent techniques are combined. These integrated approaches make the systems more proficient to do learning and reasoning activities. However, for some problems individual ITs do produce same results as integrated and it also depends on the nature of problem that is to be solved. Research in this domain is increasing year by year with new ideas and approaches. Few widely used intelligent techniques based integrated approaches are artificial neural network and data mining; genetic algorithm based fuzzy neural network; fuzzy set and gaussian dispersion model; case based reasoning, mobile agent and multi-agent; model based and rule based; knowledge based system and neural network; and knowledge based system and fuzzy theory [23]. Literature study also confirms the popularity and applicability of individual and integrated ITs for liver disease classification. From statistical techniques to data mining algorithms to neural networks, all these have been widely deployed on medical datasets for evaluating the liver sickness [22].

Liver has a vital importance to our life and is the largest internal solid organ in a human body [1]. It performs number of metabolic functions like metabolizing drugs, bile production, filtering blood, storing glucose, assisting in fat digestion, detoxifying harmful chemicals and making proteins for blood plasma [24]. Liver disease is defined as the improper functioning of liver causes illness which further leads to serious health ramifications. General causes of the disease are genetic disorders, infected eatables, immoderate consumption of ethanol, severe reaction to certain drugs, infections from bacteria and excessive fat buildup in the body [2]. Early symptoms of the disease include loss of appetite, diarrhoea, nausea, fatigue or weight loss. As time passes, the symptoms become more solemn like jaundice, hair loss, swollen abdomen, oedema, muscle wasting, memory problems, bleeding and bruising. Liver disorders are categorized in numerous types out which a few common includes viral hepatitis,

autoimmune hepatitis, primary hepatoma, neonatal hepatitis, fibrosis, cirrhosis, liver tumor, liver abscess, primary biliary cirrhosis, cholelithiasis, wilson disease, alcoholic liver disease and non-alcoholic fatty liver disease. Its presence around the world in diverse forms is increasing the mortality rate and making it a serious area of concern in medical domain. Ability of liver to function ordinarily even when partly damaged resists it from timely diagnosis and makes it more alarming because by then it has suffered significant damage. This indicates that the initiation of in time diagnosis is inevitable so that treatment can begin at right stage and in a well-organized mode. During assessment of liver sickness; selecting the features, evaluating the values, differentiating the dependent and independent predictors and finding the co-relations between various attributes stretches the decision time of physicians. To solve these impediments mentioned and to reduce the cost, time and effort needed, this research aims to develop intelligent techniques based decision-making systems for the efficient diagnosis and classification of liver disease.

1.4 Objectives of the Research

The objectives of this research work include the following:

1) To study and classify intelligent techniques applied to liver disease.
2) To develop intelligent computational models for the diagnosis of liver disease based on various clinical parameters.
3) To develop intelligent integrated approach for classifying types of liver disease.
4) To show correlation among attributes and disease, and to compare the proposed diagnostic approaches with other classification methods.

1.5 Contributions of the Thesis

The main contributions of this thesis are:

- Review of intelligent techniques applied to liver disease which would be helpful for researchers in understanding the current state-of-art in developing efficient decision-making tools. The different types of liver disorders covered in the study are: hepatitis, liver fibrosis, liver cirrhosis, liver cancer, fatty liver, general liver disorders and hepatobiliary disorders. This study also discovers the merits and demerits of medical systems developed using individual and integrated Its.
- Development of a novel integrated method based on principal component analysis and k-nearest neighbor which reduced the dimensions of data, decreased the complexity of

medical records by identifying orthogonal directions of maximum variance in the original data. This approach is superior to many bivariate statistical techniques user earlier, in that it explores the interrelationship among a set of variables caused by common factors. It also has high classification rates, good generalization, plain structure and efficient problem solving ability through feature extraction.

- Integration of dimensionality reduction techniques and classification algorithms for the prediction of primary biliary cirrhosis stages. The approach transforms features into new space which makes the classifier more efficient and there is no requirement to add a prior knowledge, even when it has very high input space dimension. The main intend was to discriminate between members of classes in training data by finding best classification function. It simultaneously maximizes the geometric margin and minimizes the classification error. Feature ranking helped in selection of vital attributes needed for input. Variation in ranking of features and correlation among features with respect to PBC stages indicated the complex nature of diagnosis process. Presence of ascites and edema were the most common and influential attributes in all four stages.
- The development of an intelligent hybrid approach for hepatitis disease diagnosis by combining enhanced k-means clustering and ensemble learning. It is a matter of prime concern to correctly predict hepatitis as this disease is also associated with other autoimmune disorders. The advantage of deploying proposed approach is that it constructs a set of hypothesis using multiple learner for solving the cases. These learning models are referred as weak learners and combination of these models creates an improved composite classification model which produces diversity in decision boundaries. It takes the advantage of model's instability by running several instances for lowering the error and the weight is recalculated to overcome the problem of misclassified gain weight. The model focuses on intricate data points which have been misclassified most by weak classifiers and shows the advantages such as high classification rates, good generalization, plain structure and efficient problem solving ability. The approach also showed capability of improving complex medical decisions through clustered data.
- Building a new intelligent model based on enhanced hierarchal clustering and random decision forest for classifying hepatobiliary disorders. The proposed model structure is more instructive than the unstructured set of clusters returned by flat clustering. It develops a sequence of nested clusters and the range is from individual clusters of single

points to all-together cluster. It resolves the problem of high bias and variance by finding average between two extremes. The integrated approach has several advantages which include enhanced prediction results through generation of smaller clusters, consistency of clusters results on different algorithms runs, precise learning, estimation of key variables, reduction in overfitting, fin computation of proximities between pairs of cases and no apriori information required about cluster numbers.

- The proposed models would be useful in hospitals and are comprehensive assisting physicians in analyzing patients' therapeutic history and training medical students.

1.6 Outline of the Thesis

This thesis is structured into eight chapters, including the first chapter which briefs the medical background of liver disease, significance of intelligent techniques, and motivation behind this research work. A brief description of each chapter's content is given.

Chapter 2 presents a comprehensive literature review on intelligent techniques applied to liver disease. ITs are divided into two categories i.e individual and integrated ITs. Individual ITs include algorithms like artificial neural network (ANN), fuzzy logic (FL) etc. Integrated ITs combine methods as artificial neural network-case-based reasoning (ANN-CBR), artificial immune system-artificial neural network-fuzzy logic (AIS-ANN-FL) etc**.** Different types of liver disease covered in this chapter are hepatitis, liver fibrosis, liver cancer, fatty liver, general liver damage, alcoholic liver damage, primary hepatoma, cholelithiasis and liver cirrhosis. The chapter identifies which individual and integrated ITs are appreciably used for what type of liver disorders with specific focus on non-invasive methods. Relative comparison in terms of usability rate of the techniques is mentioned. Performance comparison, along with their merits and demerits, of applied ITs to various liver diseases are also presented.

Chapter 3 describes the proposed scheme titled "Detection of liver disease using a novel integrated method based on principal component analysis and k-nearest neighbor". The material with its type, structure; and methods used for this work is given. Different stages of the proposed approach are illustrated using a block diagram. Firstly the principal component analysis technique is used for feature extraction. Then the performance is studied using linear discriminant analysis, diagonal linear discriminant analysis, quadratic discriminant analysis, diagonal quadratic discriminant analysis, least squares support vector machine and k-nearest

neighbor classifiers. Prediction results are compared in terms of accuracy, sensitivity, specificity, positive predictive value and negative predictive value rates.

Chapter 4 presents the proposed scheme titled "Implementation of dimensionality reduction techniques and classification algorithms for the prediction of primary biliary cirrhosis stages". Dimensionality reduction techniques are divided into two parts: feature extraction and feature selection which are implemented using principal component analysis and kullback-leibler divergence methods respectively. Detailed explanation of material and methods used for this work is given. The chapter presents two intelligent integrated models whose designs are depicted using block diagrams. In feature selection based computational methodology; first kullback–leibler divergence technique is implemented for feature ranking. Then least squares support vector machine approach is used to classify primary biliary cirrhosis stages and the prediction performance is tested using statistical parameters. In the end, critical findings about most likely association of attributes with distinct primary biliary stages are discussed. In feature extraction based computational methodology; first principal component analysis is used for extraction and then euclidean distance and nearest rule based k-nearest neighbor algorithm is used for prediction of primary biliary cirrhosis stages.

Chapter 5 discusses the proposed scheme titled "An euclidean distance function based computational model for assessing degree of liver damage". The overall structure design of the proposed approach is illustrated using a block diagram. Information about algorithms used and dataset structure is also described. The obtained results are compared with other classification algorithms using accuracy, sensitivity, specificity, positive predictive value and negative predictive value rates.

Chapter 6 explains the proposed scheme titled "An intelligent hybrid approach for hepatitis disease diagnosis: Combining enhanced k-means clustering and ensemble learning". Different stages of the approach are illustrated using a block diagram. Detailed explanation of material and methods used for this work is given. Experimental results proved the superiority of proposed integrated method in comparison to other classification models. Statistical parameters computed for evaluating the system performance are also defined in the chapter.

Chapter 7 presents the proposed scheme titled "A new intelligent model based on enhanced hierarchal clustering and random decision forest for classifying hepatobiliary disorders". Different stages of the methodology are illustrated using a block diagram.

Experimental results of the intelligent integrated model are obtained in terms of accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. The chapter also discusses the improved performance of proposed approach in comparison to other individual and integrated classifiers build in the study, and classification methods mentioned in the literature.

Chapter 8 draws the conclusion of thesis with a summary of results. It also presents the future scope of work. The list of publications related to the research work is mentioned at the end of thesis.

Chapter 2

Literature Review

This chapter is organized as follows: Section 2.1 introduces the need of literature review and significance of intelligent techniques in liver disease. Section 2.2 presents the survey on individual intelligent techniques which incorporate ANN, DM, FL and EDC. Section 2.3 covers integrated intelligent techniques include ANN-CBR, ANN-DM, ANN-FL, AIS-FL, ANN-GA, AIS-GA, ANN-PSO, CBR-DM, CBR-GA, DM-GA, DM-FL, FL-GA, AIS-ANN-FL, ANN-CBR-RBR, AIS-DM-FL, ANN-DM-FL, ANN-DM-GA, ANN-GA-RBR, CBR-GA-PSO, DM-FL-GA and CBR-DM-FL-GA. Section 2.4 presents the observations. Finally, conclusions are drawn in Section 2.5.

2.1 Introduction

Intelligent techniques based frameworks have played a vital role in liver disease diagnosis. From statistical techniques to data mining algorithms to neural networks, all these have been widely deployed on liver examination data for evaluating the sickness. Due to increasing vagueness and complexities in the datasets, deriving intelligible information becomes a major challenge for clinicians. This challenge could lead to imprecise assessment of the disease, which would further channelize inaccurate treatment to patients. So to avoid these uncertainties up to a feasible extent, medical professionals refer the intelligent decision-making systems for a second thought on the interpretation of patient's data. Like for other health complications, intelligent techniques have also been widely applied to diagnose and classify liver disease. Liver disease is a serious area of concern in the universal set of medicine and is becoming the leading cause of mortality around the globe. Implementing these ITs in liver disorders is acting as a catalyst in overcoming the overheads and problems faced by doctors. ITs effectively prevail over the inadequacies, help in obtaining better accuracies and make the systems adaptable. ITs decrease the probability of occurrence of medical errors, and reduces the cost, time and effort needed. Going into more specific discussion, intelligent techniques applied to liver disorders discussed in this chapter are as follows: Artificial neural network (ANN), data mining (DM), fuzzy logic (FL), evolutionary data classification (EDC), artificial neural network-case-based reasoning (ANN-CBR), artificial neural network-data mining (ANN-DM), artificial neural network-fuzzy logic

(ANN-FL), artificial immune system-fuzzy logic (AIS-FL), artificial neural network-genetic algorithm (ANN-GA), artificial immune system-genetic algorithm (AIS-GA), artificial neural network-particle swarm optimization (ANN-PSO), case-based reasoning-data mining (CBR-DM), case-based reasoning-genetic algorithm (CBR-GA), data mining-genetic algorithm (DM-GA), data mining-fuzzy logic (DM-FL), fuzzy logic-genetic algorithm (FL-GA), artificial immune system-artificial neural network-fuzzy logic (AIS-ANN-FL), artificial neural network-case-based reasoning-rule-based reasoning (ANN-CBR-RBR), artificial immune system-data mining-fuzzy logic (AIS-DM-FL), artificial neural network-data mining-fuzzy logic (ANN-DM-FL), artificial neural network-data mining-genetic algorithm (ANN-DM-GA), artificial neural network-genetic algorithm-rule-based reasoning (ANN-GA-RBR), case-based reasoning-genetic algorithm-particle swarm optimisation (CBR-GA-PSO), data mining-fuzzy logic-genetic algorithm (DM-FLGA) and case-based reasoning-data mining-fuzzy logic-genetic algorithm (CBR-DM-FL-GA). This chapter has made a contribution to medical field by presenting a study on intelligent techniques applied to liver disorders. To the best of our knowledge, not a single attempt had been made to write any review on liver disease for the last 40 years (1976–2016). Numerous authors have written literature review sections but no complete review has found so far. This study would be helpful for researchers in developing efficient decision-making tools as one need to be well acquainted with the applicability of ITs in liver disorders. Different types of liver disease covered are hepatitis, liver fibrosis, liver cancer, fatty liver, general liver damage, alcoholic liver damage, primary hepatoma, cholelithiasis and liver cirrhosis. This chapter has figured out which ITs are widely used and vice-versa, which ITs outperformed others in comparison and what are the attributes taken for experimentation. It could also be possible that researchers used some methods that are not considered for this survey but we have mentioned those methods, in table, wherever possible. The study also discovers the merits and demerits (if any) of medical systems developed using individual and integrated ITs.

2.2 Individual Intelligent techniques

This section briefly introduces individual intelligent techniques applied to liver disease. The survey is detailed in tabular form containing following information: author name, year of publication, attributes, intelligent techniques and other methods used, and result and application. Based on the literature survey, merits and demerits (if any) of IT based systems are also stated. Individual ITs discussed in this segment are ANN, DM, FL and EDC.

2.2.1 Artificial Neural Networks (ANNs)

Artificial neural networks are a simulated view of human brain that is composed of artificial neurons or nodes. These neurons are made up of highly interconnected and interacting processing units [3]. The first design of ANNs was given by physiologists, McCulloch and Pitts in the year 1943. The basic structure of ANN consists of input, hidden and output layers which collectively work as a neuron of a human brain. Neurons in ANN communicate with each other with the help of impulses. These impulses could be of dual nature such as electrical or chemical. It works by acquiring raw data from the outer world for generalizing the knowledge. ANNs have an immense ability to learn and derive meaning from intricate and imprecise data. It does not require any preceding knowledge of a problem [25]. ANN has impactful applications in various fields like pattern recognition, time series prediction, data processing, robotics, regression analysis etc. Apart from these domains ANN has also spread its reach in medicine too. It is emerged as one of the most popular tool and provides promising results in medical data analysis [26]. In medicine, ANN helps in diagnosis, radiology, medical image analysis etc. Limitations of ANN are: it requires lots of knowledge intake, ANN based systems take lots of time to get fully trained and at times it is difficult to find adequate solution to a problem. Apart from this, choosing an appropriate knowledge set is yet again a major challenge.

ANNs have fascinated many researchers and have shown remarkable performance when applied to liver disorders. ANN based systems are reliable, robust, more accurate, predictive, computationally simple, non invasive and inexpensive [4], [27]–[31]. ANN speeds up the learning process and solves fast size-growing problem [32]. Levenberg marquardt training algorithm of multilayer perceptron (MLP) network employing back propagation shows fair prediction and obtains lower mean square errors [29]. Implementation of MLP networks trained by resilient back propagation algorithm is good in improving classification accuracy of small classes [30]. Elizondo et al. (2012) proposed a method which detects differences in the complexity of classification problem [33]. Sun et al. (2005) deployed fast discrete wavelet transform for decreasing time consumption in computation [34]. Some limitations of ANN based systems are: it is difficult to explain complex classification process as rules [32], classification accuracy for MLP is low [27], self organizing map (SOM) is unreliable to diagnose hepatitis virus [35], implementation of pyramid neural network consumes a bit longer time in processing [34].

Hayashi et al. (2000) stated that overall accuracy rates obtained from NeuroLinear and NeuroRule are higher than those of linear discriminant analysis (LDA) and fuzzy neural networks (FNN) [36]. Ozyilmaz and Yildirim (2003) found the accuracy of conic section function neural network (CSFNN) higher than C4.5 decision tree, naive bayes classifier, bayesian network with naive dependence and feature selection (BNNF) [27]. Ansari et al. (2011) asserted that supervised model performs better as compared to unsupervised one [35]. Perez et al. (2012) proposed an associative memory based classifier (AMBC) method that achieves highest classification accuracy among methods such as AdaBoostM1, Bagging, BayesNet, Logistic [37]. Generalized regression neural network (GRNN) performed better than feedforward back propagation neural network (FFNN) [35]. Revesz and Triplet (2010) compared classification of integration and data integration methods, both used support vector machine (SVM) linear classifier, and found that the former is more accurate than latter in case of missing values in data [38]. Two-level neural network (2-LNN) [39] method achieved higher predictive accuracy than fuzzy neural network (FNN) and fuzzy multilayer perceptron (FMLP). ANN [28] attained better classification accuracy than decision tree and multivariate logistic regression analysis (MLRA); performed better than MLRA in simulating nonlinear relation between fibrosis grades and biomarkers. Radial basis functions (RBF) network [40] outperformed all networks including GRNN, learning vector quantization network (LVQ), PNN and SVM, except for the one class (Hepatitis B) in which the probabilistic neural networks (PNN) performed better. SVM-SA based system [41] achieved better accuracy than methods like C4.5, naive bayes (NB), LDA, LVQ, GA-SVM, MLP, GRNN, and MLP with BP (back propagation). The survey on applicability of ANNs for liver disorders is listed in Table 2.1.

Table 2.1: Details of ANN based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and application**
Hamamoto et al. (1995) [42]	Preoperative aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, total bilirubin of the serum, hepaplastine	Perceptron-type neural network, linear regression method, supervised learning based on back-propagation method	Prediction of early prognosis of hepatectomized patient with hepatocellular carcinoma (liver cancer). Accuracy:100%

	test, $ICGR_{15}$, total liver volume, residual liver volume and number of platelet		
Hayashi et al. (2000) [36]	Hepatobiliary disorders dataset	Standard feedforward network, line search algorithm, quasi newton algorithm, BFGS method, NeuroLinear and NeuroRule rule extraction techniques	Diagnosis of hepatobiliary disorders. Accuracy: NeuroRule - 88.3% Neuro Linear - 90.2%
Hayashi and Setiono (2002) [39]	Hepatobiliary disorders dataset	Standard feedforward network with a single hidden layer	Diagnosis of hepatobiliary disorders. Accuracy of 2-LNN: 83.47% (using best choice criterion) and 91.41% (using second best choice criterion)
Ozyilmaz and Yildirim (2003) [27]	Hepatitis dataset	MLP trained with standard back propagation algorithm, RBF trained with OLS algorithm, CSFNN combined MLP and RBF, gaussian bell function	Diagnosis of hepatitis disease. Accuracy: CSFNN - 90% RBF - 85% MLP - 81.375%
Lee et al. (2005) [32]	Contour of the liver cyst, contrast between liver tissues, gray levels of the liver tissues	BP-CMAC neural network (integrated with BP and CMAC)	Classification of liver disorders (liver cyst, hepatoma and cavernous hemagioma).

			Accuracy: 87%
Mala and Sadasivam (2005) [43]	Computerized tomography images	Probabilistic neural network. Wavelet based texture analysis: orthogonal wavelet transform, mean, standard deviation, contrast, entropy, homogeneity and angular second moment	Classification of diffused liver disorders (fatty liver and liver cirrhosis). Accuracy: 95% Sensitivity: 96% Specificity: 94%
Sun et al. (2005) [34]	Ultrasonographic images of cirrhosis	Pyramid neural network trained using ultrasonographic images of cirrhosis and using data judged by clinicians, fast discrete wavelet transform, steepest descent method	Diagnose the types of cirrhosis disease.
Azaid et al. (2006) [31]	Ultrasound images: mean gray level, variance of gray levels, skewness of gray level distribution, kurtosis	Multi layer back propagation neural network trained on features (mean gray level, variance of gray level, skewness of gray level and kurtosis), quantitative tissue characterization technique, square	Classify liver disorders as fatty liver, liver cirrhosis, liver cancer. Accuracy: 96.125%

		shaped region technique	
Revett et al. (2006) [44]	Case number, days since registration, drug, age at initial registration, sex, days between study enrollment and a visit, presence of ascites, presence of hepatomegaly, presence of spiders, presence of edema, serum bilirubin, serum cholesterol, albumin, alkaline phosphatase, sgot, platelets, prothrombin time, histologic stages of disease	Probabilistic neural network, approach based on bayes formula, taylor's polynomial approximation. Rough Sets: rosetta implementation, entropy preserving or MDL (minimal description length) algorithm, equivalence classes	Mining a primary biliary cirrhosis dataset. Accuracy: 87%
Icer et al. (2006) [29]	Power spectral densities (PSD) of doppler signals	Feed forward multi layer perceptron network, sigmoid transfer functions, training algorithms adopted were resilient propagation algorithm (RP), scaled conjugate gradient algorithm (SCG) and levenberg-marquardt algorithm (LM) employing	To determinate cirrhosis disease with power spectral densities of portal venous doppler signals. Accuracy, sensitivity and specificity was 100% with levenberg-marquardt training algorithm

		backpropagation, power spectral densities (PSD) of portal venous doppler signals, short time fourier transform (STFT) method	
Autio et al. (2007) [30]	Liver disorders dataset	Multilayer perceptron networks trained with resilient backpropagation algorithm, logarithmic sigmoid function, root mean square formula, least gradient technique, tenfold cross validation	Classification of liver disorders as sick and healthy. Accuracy: 71 %
Dong et al. (2008) [45]	Liver disorders dataset	Support vector machines, tenfold cross validation	To calculate optimal value of cost parameter in order to minimize classification error. Accuracy: 68.12%
Su and Yang (2008) [46]	Liver disease dataset collected from department of health examination (chang gung memorial hospital, tao-yuan, taiwan)	Support vector machine model, polynomial kernel, gaussian radius base function kernel and combined kernel functions, L-J method for feature selection	Classification of liver disorders. Accuracy: 77% (with 100% features)

Rouhani and Haghighi (2009) [40]	Sex, age, ALK, AST, ALT, Bi, T, D, G.G.T, HBSAg, ALB, LHD, PT, FBS, CHO and HCVAb	RBF networks: two-layer structure, linear activation function, gaussian function GRNN: radial basis layer and a special linear layer PNN: structure was alike RBF networks, gaussian distribution, competitive transfer function LVQ networks: competitive layer and a linear layer	Diagnosis of hepatitis disease. Accuracy (RBF): 96.4%
Uttreshwar and Ghatol (2009) [47]	Hepatitis B surface antigen (HBsAg), hepatitis B surface antibody (HBsAb), hepatitis B e-antigen (HBeAg), hepatitis B DNA	Generalized regression neural networks, kernel-based approximation, logical inference, IF-Then rules	Diagnosis of hepatitis B. Accuracy: 86.3237%
Bucak and Baki (2010) [4]	AST, ALT, AST/ALT, albumin, protein, platelet, and prothrombin time	CMAC neural network, supervised learning, quantization, least mean square (LMS)	Diagnosis of liver disorders (hepatitis B, hepatitis C, cirrhosis A, cirrhosis B and C). Accuracy: 100%
Hashem et al. (2010) [28]	Routine work tests: platelets count, hemoglobin, WBCs, RBCs, alanine aminotransferase, aspartate	ANN: simulate nonlinear relation between fibrosis grades and biomarkers Analysis of variance	Prediction of the degree of liver fibrosis (predict the hepatic fibrosis extent in patients with HCV). Accuracy: 93.7% Sensitivity: 92.5%

	aminotransferase, alkaline phosphatase, serum albumin, total bilirubin, prothrombin concentration, alfa-fetoprotein, thyroid stimulating hormone, creatinine, urea, and blood glucose. Fibrotic markers: matrix metalloproteinase-1, metalloproteinase-2, hyaluronic acid, tissue inhibitor of metalloproteinase-1, tissue growth factor beta1, α2-macroglobulin, haptoglobin, apolipoprotineA1	(ANOVA), tukey-kramer and bonferroni multiple comparison tests, sequential R-square measure, box plots	Specificity: 94.8% Area under ROC curve: 0.974
Revesz and Triplet (2010) [38]	Case number, days between registration and earliest of death, transplantation or study; age in days, gender, ascites present, hepatomegaly present, spiders present, edema, serum bilirubin, serum cholesterol, albumin, urine copper, alkaline phosphatase, sgot, triglicerides, platelets,	linear kernel support vector machine, r-square measure, box plots	Liver disorder classification (primary biliary cirrhosis)

	prothrombin time in seconds, status, drug, histologic stage of the disease		
Ansari et al. (2011) [35]	Hepatitis dataset	FFNN: levenberg-marquardt back propagation algorithm, mean square error (MSE) formula. GRNN: kernel function, radial basis transfer function, linear transfer function, euclidean distance weight function. SOM: competitive learning approach, euclidean distance, link distance function.	Diagnosis of hepatitis virus. Accuracy: FFNN-91.3%, GRNN-92%
Arsene and Lisboa (2012) [48]	Time, triglicerides, SGOT, serum cholesterol, alkaline phosphatase, ascites, platelets, urine copper, spiders, bilirubin, albumin, age, gender, presence of edema, prothrombin time, hepatomegally, histologic stage of	Bayesian neural network, partial logistic artificial neural network (PLANN) and automatic relevance determination (ARD), bayesian regularization framework, hessian matrix of the total	Medical survival analysis of primary biliary cirrhosis (PBC)

	disease, drug	error function, local and a global compensation mechanism	
Perez et al. (2012) [37]	Liver disorders dataset and Hepatitis dataset:	Associative memory based classifier (AMBC): learning phase, learning reinforcement phase and classification phase, IntegerToVector operator, tenfold, holdout and leave-one-out cross validation methods	Diagnosis of liver disorders. Classification accuracy using 50-50 training-test split: BUPA-65.40% Hepatitis-83.76% Classification accuracy using 70-30 training-test split: BUPA-59.593% Hepatitis-84.86% Classification accuracy using 10 fold cross validation: BUPA-65.50% Hepatitis-85.16% Classification accuracy using leave-one-out cross validation: BUPA-60.57% Hepatitis-85.16%
Elizondo et al. (2012) [33]	Hepatitis dataset	Recursive deterministic perceptron (RDP) neural network, simplex method for testing linear separability,	Quantifying the level of complexity of classification datasets

		ANOVA analysis	
Sartakhti et al. (2012) [41]	Hepatitis dataset	RBF kernel based support vector machine, simulated annealing, k-fold cross validation	Hepatitis disease diagnosis. Accuracy: 96.25%
Babu and Suresh (2013) [49]	Liver disorders dataset	PBL-McRBFN: cognitive component and meta-cognitive component, gaussian activation function, sample delete strategy, neuron growth strategy, parameter update strategy, sample reserve strategy	Classification performance on liver disorders dataset. Accuracy: 72.63%
Jeon et al. (2013) [50]	Ultrasound images: characteristics of lesions including internal echo, morphology, edge, echogenicity and posterior echo enhancement	Multiple ROI selection, support vector machine, feature-level fusion method to combine features, tenfold cross validation	Focal liver lesion classification. Accuracy for classification of cysts and hemangiomas: 93.77% Accuracy for classification of cysts and malignancies: 92.13% Accuracy for classification of hemangiomas and malignancies: 80%

CMAC: Cerebellar Model Articulation Controller, PBL: Projection Based Learning, McRBFN: Meta-cognitive Radial Basis Function Network, ROI: Region of Interest. Liver disorders dataset (LDD) attributes are: mcv (mean corpuscular volume), alkphos (alkaline phosphotase), sgpt (alamine aminotransferase), sgot (aspartate aminotransferase), gammagt (gamma-glutamyl transpeptidase), drinks (number of half-pint equivalents of alcoholic

beverages drunk per day). Hepatitis dataset (HD) attributes are: age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology. Hepatobiliary disorders dataset (HDD) attributes are: Glutamic Oxalacetic Transaminate (GOT, Karmen unit), Glutamic Pyruvic Transaminase (GPT, Karmen Unit), Lactate Dehydrase (LDH, iu/l), Gamma Glutamyl Transpeptidase (GGT, mu/ml), Blood Urea Nitrogen (BUN, mg/dl), Mean Corpuscular Volume of red blood cell (MCV, fl), Mean Corpuscular Haemoglobin (MCH, pg), Total Bilirubin (TBil, mg/dl) and Creatinine (CRTNN, mg/dl).

2.2.2 Data Mining (DM)

Data mining is a process of identifying hidden relationships and discovering new knowledge from large datasets. The term "Data Mining" was originated in the year 1990, but work on this had started a bit earlier. DM helps in processes such as classification, clustering, regression and summarization which generate hidden facts from historical data. Some issues in DM which needs to be taken care of are: quality of data to mine, the extent up to which data needs to be cleaned, and interoperability (data from heterogeneous sources needs to combined and analyzed). DM based system has the capability of replacing liver biopsy in liver disorders diagnosis [51]. Yan et al. (2008) C4.5 decision tree based proposed method can be efficiently integrated with algorithm like boosting to enhance prediction [52]**.** Eastwood and Gabrys (2012) mentioned advantages of a single tree classifier includes its simple structure, small memory requirement and quick calculation of predictions [53]. Kohara et al. (2010) has done something very interesting and out of the box by proving the feasibility to diagnose liver cirrhosis using PCA based statistical shape model of the liver [54].

In comparison, Yan et al. (2008) proposed C4.5 algorithm has better classification rate than methods like ID3 decision tree, RBF NN, BayesNet and logistic [52]. Based on the results achieved, Yan et al. (2008) claimed that if a patient has gotten cirrhosis, he must have symptoms like lassitude and fatigue, chill and cold limbs, tarnish complexion, yellow eyes or yellow body or yellow urine. Floares (2009) developed a C5.0 decision tree and boosting based system which has outperformed other methods such as support vector machines, bayesian networks, neural networks of various types and architectures, and classification and regression trees [51]. The survey on applicability of data mining techniques for liver disorders is listed in Table 2.2.

Table 2.2: Details of DM based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Yan et al. (2008) [52]	Lassitude and fatigue, chill and cold limbs, tarnish complexion, yellow eyes or yellow body or yellow urine	C4.5 decision tree, tenfold cross validation method	To analyze relationship between child-pugh degree and examinations of traditional chinese medicine based on liver cirrhosis. Accuracy: 85.67 % (for child pugh A)
Floares (2009) [51]	Age, aspartate aminotransferase, gamma-glutamyl-transpeptidase, cholesterol, triglycerides, thickness of the gallbladder wall, spleen area and perimeter, left lobe and caudate lobe diameter, liver homogeneity, posterior attenuation of the ultrasound, liver capsule regularity, spleen longitudinal diameter, the maximum subcutaneous fat, perirenal fat	C5.0 decision tree and adaptive boosting	Liver disorders classification (chronic hepatitis C and B). Accuracy: 100%
Kohara et al. (2010) [54]	Components of shape feature vector	Marching cube algorithm, chui method, principal	Diagnosis of liver cirrhosis

		component analysis	
Luo et al. (2011) [55]	Jaundice, poor appetite, fatigue, yellow urine and hypochondriac pain	Cluster analysis: DBScan algorithm Association rules: Apriori algorithm	Preventing and treating viral hepatitis
Eastwood and Gabrys (2012) [53]	Liver disorders dataset	Standard decision tree induction, resampling (bootstrapping), model level combination method, pessimistic pruning and error-based pruning, tenfold cross validation	Proposed a pruning criteria (Liver disorders database for the empirical investigation of proposed method).
Jen et al. (2012) [56]	Systolic pressure, diastolic pressure, glutamate-pyruvate transaminase, alpha-fetoprotein	K-nearest neighbor, linear discriminant analysis with sequential forward selection (a bottom-up search procedure)	Used risk factors of chronic diseases (disease of the liver) to build early warning criteria. Accuracy: 82.65%

2.2.3 Fuzzy Logic (FL)

FL based models have been developed and utilized by numerous researchers for assessing liver disease. The concept of fuzzy logic was introduced by Lotfi A. Zadeh in the year 1965 and was applied in medical systems approximately after 20 years. FL employs linguistic rules in the form of IF-Then statements. FL deals with uncertainty and assists computer in interpreting statements which consists of intermediate constructs. For example, if glass is half full, then pour some water. The semantic of this statement does not correspond to any true value, either true/false. FL based systems are faster, liable, cheaper [57], robust, flexible, customizable, interpretable and easy to train [58]. FL based systems also has flexible initialization, fast convergence and robust segmentation [1]. FL based systems efficiently deals with uncertainty, ambiguous information and imprecise data [59]. Ming et al. (2011)

fuzzy based framework uses global k-means algorithm to determine actual number of cluster needed for different datasets and fast global k-means to improve computation time taken by global k-means algorithm [60]. This model was based on enhanced supervised fuzzy clustering algorithm which effectively handles small size data that is noisy and atypical. Sometimes fuzzy based systems also require more simulation and fine tuning.

Based on the comparisons, Badawi et al. (1999) proposed fuzzy based classification attains higher sensitivity than neural network classification, and higher sensitivity and specificity than statistical classification techniques [61]. Obot and Udoh (2011) stated that FL ability to work from approximate reasoning and finding precise solution makes it superior to methods such as ANN, RBR and CBR [59]. Neshat et al. (2008) proposed fuzzy system obtains higher accuracy than other traditional diagnostic systems such as RULES-4, C4.5, Naive Bayes, BNND, BNNF, SVM with GP, SSVM, RSVM, MLP, PNN, GRNN, RBF, AIRS and FW-AIRS [57]. Gadaras and Mikhailov (2009) proposed fuzzy classification framework achievs higher accuracy than other techniques mentioned in literature like FBP-NN, BZ, GF-SVM and NF-BSP [58]. Ming et al. (2011) proposed a deterministic and autonomous algorithm (enhanced supervised fuzzy clustering) which attains higher mean accuracy than supervised fuzzy clustering method [60]. Luukka (2011) proposed fuzzy bean based classifier obtains higher accuracy than classifiers like CN2, MLP, DIMLP and SIM [62]. The survey on applicability of FL for liver disorders is listed in Table 2.3.

Table 2.3: Details of FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Badawi et al. (1999) [61]	Mean gray level, contrast, angular second moment, entropy, correlation, attenuation and speckle separation	Fuzzy rules, MIN compositional rule of inference, bell membership function	Differentiate diffuse liver disorders. Results of fuzzy rule-based classification: Specificity: 92% Sensitivity for liver cirrhosis: 94% Sensitivity for fatty liver: 96%
Neshat et al.	Liver disorders dataset	Fuzzy rules,	Liver disorders diagnosis

(2008) [57]		triangular or trapezoidal fuzzifier, center of gravity defuzzifier formula	(healthy and unhealthy liver). Accuracy: 91%
Gadaras and Mikhailov (2009) [58]	Liver disorders dataset	Fuzzy rules, min-max method, trapezoid membership function	Classification performance on liver disorders dataset. Accuracy: 89.9%
Luukka (2011) [62]	Liver disorders dataset	Fuzzy beans, bocklisch membership function, differential evolution algorithm	Liver disorders diagnosis. Accuracy: 73.9%
Ming et al. (2011) [60]	Hepatobiliary disorders dataset	Enhanced supervised fuzzy clustering algorithm, unsupervised gath-geva algorithm	Liver disorders classification (alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis). Accuracy: 58.78%
Obot and Udoh (2011) [59]	Nausea, vomiting, fever, body weakness, loss of appetite, diarrhea, itching, convulsion, stupor, headache, tremors, skin discoloration, eye discoloration, liver tenderness, bile in urine, jaundice	Fuzzy rules, max-min method, centre of gravity (CoG) method, fuzzified with membership functions	Diagnosis of hepatitis
Li et al. (2012) [1]	Contrast-enhanced computed tomography images	Unsupervised fuzzy clustering, fuzzy c-means	Semi-automatic liver tumor segmentation

2.2.4 Evolutionary Data Classification (EDC)

EDC is the classification of data through evolutionary algorithms. In this approach prediction in data is done with the help of evolutionary learning. The most widely used evolutionary approach is genetic algorithm. The study of genetic algorithm was proposed in 1975 by John Holland at university of michigan, united states [63]. GA is a branch of evolutionary algorithms which imitates the process of natural evolution and survival of the fittest [64]. It finds the optimum solution from the set of candidate solutions. It uses a fixed-length chromosome structure and is aimed at solving optimization or search problems. The basic requirements of GA are: a genetic representation of solution set, and a fitness function to test and evaluate the solution set. Primary genetic operators used by GA are selection, crossover, and mutation. It has several limitations like implementation of fitness function to evaluate the solution set is quite expensive, shows less efficiency as the complexity of the problem increases and does not operate well on dynamic datasets. Despite this, usage of GA for optimizing parameters makes the diagnostic systems robust and invariant [65]. Falco (2013) proposed a tool that extracts knowledge in the form of IF-Then rules from databases [66]. This is simple, faster, robust, reliable and easy to implement. It also helps users in medical diagnosis and gives explanation of evidences on why a patient is suffering from a specific disease.

In comparison, Tan et al. (2003) proposed two-phase hybrid evolutionary classification technique performed better than methods such as C4.5 (decision tree program), PART (rule-learning scheme) and is comparable to naive bayes (utilizes the bayesian techniques) [65]. Zhang and Rockett (2011) proposed feature extraction method proves its superiority to competitive methods such as RBF, logistic (modified multinomial logistic regression model), nearest-neighbour-like algorithm, bayes network classifier using K2 learning algorithm, instance-based learning algorithm, ADTree (the alternating decision tree learning algorithm), sequential minimal optimisation (SMO) algorithm and C4.5 decision tree algorithm [64]. This method also records lowest mean error. Wu et al. (2012) presented a GA-based feature selection algorithm which selects a better feature subset than serial feature combination and serial feature fusion schemes [67]. This algorithm performs better, in selecting feature subsets, than NMIFS (normalized mutual information feature selection) and GAMIFS (a hybrid filter/wrapper method called GAMIF). Falco (2013) proposed tool is superior to bayes net, naive bayes, IB 1, FLR (fuzzy lattice reasoning), VFI (voting feature interval), OneR, Part, and inferior to MLP, RBF, KStar, AdaBoostM1, bagging, ridor (ripple down rule), J48,

NBTree [66]. It also requires lowest number of rules in comparison to other rule-based classification methods (Part, OneR and Ridor). Inspite of these excellent features, this tool has a limitation of not taking uncertainty into account. The survey on applicability of GA for liver disorders is listed in Table 2.4.

Table 2.4: Details of GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Tan et al. (2003) [65]	Hepatitis dataset	Tournament selection scheme, genetic programming tree-based chromosome representation, ramped-half-and-half approach, fixed-length real-coding chromosome structure, standard tree-based crossover and mutation operators, standard single-point crossover, covering algorithm, knowledge presented as multiple IF-Then rules, michigan coding approach, pittsburgh coding approach, pittsburgh-like approach, paired t-test	Predict whether a patient with hepatitis will live or die. Average Accuracy: 83.92% Best Accuracy: 94.34%
Zhang and	Liver disorders dataset	Binary tournament	Classification performance

Rockett (2011) [64]		selection, depth-fair operator, roulette wheel selection, non-destructive, depth dependent crossover and mutation operators, minimization of vectors, three-dimensional fitness vector compromising tree complexity	on liver disorders dataset.
Wu et al. (2012) [67]	Ultrasonic liver image dataset	Two-point crossover and mutation, roulette wheel selection scheme, k-nearest neighbor method, threefold cross validation	Ultrasonic liver tissue characterization (cirrhosis, hepatoma, and normal). Accuracy: 96.62 %
Falco (2013) [66]	Liver disorders dataset	Differential evolution method, tenfold cross validation mechanism	Automatic classification of items in medical databases. Accuracy in case of liver disorders dataset: 64.74% Specificity: 45.08% Sensitivity: 79.84% ROC curve area: 62.46

2.3 Integrated Intelligent Techniques

This section presents the review results of integrated intelligent techniques applied to liver disorders. The study is detailed in tabular form containing following information: author name, year of publication, attributes, intelligent techniques and other methods used, and result and application. Integrated ITs combine methods in one of the two ways: either the techniques are applied sequentially in which one technique is used to accomplish a specific task that is

followed by second technique and so on, or all the techniques are applied simultaneously. For example, sometimes integration of both ANN and CBR is used to identify the existence of liver disorder, whereas in other integration ANN is used to identify the existence of liver disorder and CBR is used to find the types of liver disorder. It could also be possible that researchers used some methods that are not considered for this survey but we have mentioned those methods, in the table, wherever possible. This section also enlightens the benefits of integrated ITs based systems when used for liver disorders. Integrated intelligent techniques focused in this segment are ANN-CBR, ANN-DM, ANN-FL, AIS-FL, ANN-GA, AIS-GA, ANN-PSO, CBR-DM, CBR-GA, DM-GA, DM-FL, FL-GA, AIS-ANN-FL, ANN-CBR-RBR, AIS-DM-FL, ANN-DM-FL, ANN-DM-GA, ANN-GA-RBR, CBR-GA-PSO, DM-FL-GA and CBR-DM-FL-GA.

2.3.1 ANN-CBR

ANN-CBR methodology was used by Lin and Chuang (2010) where ANN is deployed to examine the existence of liver disorder and AHP-weighted CBR is used to discover the types of liver disorder [3]; and by Chuang (2011) where ANN-CBR integration is deployed to obtain enhanced accuracy in diagnosis [2]. The integration of ANN-CBR makes the diagnosis more accurate and comprehensive [2], decreases the occurrence of false diagnosis and avoids postponement of treatment [3]. Chuang (2011) made it evident that proposed ANN-CBR model achieves better diagnostic accuracy than BPN (back-propagation neural network), CART (classification and regression tree), DA (discriminatory analysis), LR (logistic regression), CBR, LR-CBR, DA-CBR, and CART-CBR. Lin and Chuang (2010) used AHP-weighted CBR instead of CBR because it reduces diagnostic errors, accelerates the medical treatment and most importantly has obtained better accuracy. AHP allocated weights to the attributes. One appealing fact in this study (Lin and Chuang, 2010) was identifying types of liver disorders as most of the literature work had not moved beyond diagnosis. The survey on applicability of ANN-CBR for liver disorders is listed in Table 2.5.

Table 2.5: Details of ANN-CBR based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Lin and Chuang (2010) [3]	Hepatitis test: HBsAg, HBeAg, Anti-HBs, Anti-HBe, Anti-	ANN: BPN trained with gradient steepest descent	To examine the existence of liver disorder and to determine the types of liver

	HBc Anti-HCV Liver function test: AST (SGOT), ALT, T-Bil, ALB, ALP, r-GT. Tumor marker: α-AFP Basic information: Gender, marriage, blood type, age, education, occupation. Lifestyle habit: Tattoo, smoking, chewing betel-nut, alcohol. Lifestyle: Fatigue, sleep, nap, exercise, breakfast habit, vegetables, fruits, food date mark, food composition, low-salt, low-sugar. Health condition: Healthy status, weight, response to physical discomfort, healthy examination, acupuncture blood donation	algorithm. CBR: retrieve most similar case, vector of features, case indexing, case retrieval, assigning weights to attributes, nearest neighbor method. AHP: structure decision hierarchy, pair wise comparisons, initiate prioritization evaluated consistency, means of a consistency ratio, compute relative weights, geometric mean. Fivefold cross validation	disorder. Accuracy: ANN (diagnosis of liver disorders): 98.04% AHP-weighted CBR (discovers the types of liver disorders): 94.57% Types of liver disorders: 90.2% for chronic hepatitis, 19.6% for liver cirrhosis, 60.2% for B hepatitis and 10% for alcohol hepatitis
Chuang (2011) [2]	Hepatitis test: HBsAg, HBeAg, Anti-HBs, Anti-HBe, Anti-HBc Anti-HCV. Liver function test:	ANN: BPN implemented using NeuroShell 2.0, gradient steepest descent training	Liver disorders diagnosis Accuracy: BPN-CBR: Accuracy: 95% Sensitivity: 98%

	AST (SGOT), ALT, T-Bil, ALB, ALP, r-GT. Tumor marker: α-AFP Basic information: Gender, marriage, blood type, age, education, occupation. Lifestyle habit: Tattoo, smoking, chewing betel-nut, alcohol. Lifestyle: Fatigue, sleep, nap, exercise, breakfast habit, vegetables, fruits, food date mark, food composition, low-salt, low-sugar. Health condition: Healthy status, weight, response to physical discomfort, healthy examination, acupuncture blood donation	algorithm. CBR: euclidean distance to extract similar cases, nearest neighbor method. Tenfold cross validation, sampling using the bernoli method	Specificity:94% AUC: 96% BPN: Accuracy: 93%, Sensitivity: 91% Specificity: 96% AUC: 93% CBR: Accuracy: 89% Sensitivity: 90% Specificity: 88%, AUC: 89%

2.3.2 ANN-DM

ANN-DM methodology was used by Bologna (2003) where DM is used for extraction of rules and ANN is used for classification [68]; and by Calisir and Dogantekin (2011) where DM is used for feature extraction and feature reduction, and ANN is used for classification [69]. ANN-DM based systems in medical domain have reliability, more accuracy, small-sample problem solving ability, correct recognition rates, simple structure and good

generalization [68]–[70]. Calisir and Dogantekin (2011) proposed model achieves higher accuracy than methods such as Weighted9NN, 18NN, ASI, MLP+BP (Tooldiag), LDA, MLP, RBF (Tooldiag), 1NN, RBF, FS-AIRS, 15NN, FSM with rotations, FSM without rotations, MLP with BP, QDA, Naive Bayes, Fisher discriminant analysis (FDA), LVQ, GRNN, ASR, IncNet, CART, PCA-AIRS, and LFC. Bologna (2003) proposed DIMLP model is appreciably more accurate than CN2 induction algorithm on most of the problems. The survey on applicability of ANN-DM for liver disorders is listed in Table 2.6.

Table 2.6: Details of ANN-DM based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Bologna (2003) [68]	Hepatitis dataset and Liver disorders dataset	ANN: discretized interpretable multi-layer perceptron (DIMLP) model, staircase activation function, sum squared error (SSE) function, gradient was determined using sigmoid functions, trained by back-propagation with default parameters, bagging and arcing methods based on resampling techniques, relevance hyperplane criterion. DM: C4.5 decision trees, IF-Then rules. Tenfold cross validation, t-statistic test, two tailed test	Diagnosis of liver disorders. Average predictive accuracy: Hepatitis: 79.1% Liver disorders: 70.15%

Calisir and Dogantekin (2011) [69]	Hepatitis dataset	ANN: least squares support vector machine, maximum euclidean distance, parameters includes width of gaussian kernels and regularization factor. DM: principle component analysis	Diagnosis of hepatitis disease. Accuracy: 96.12%
Hashem et al. (2012) [70]	HA, TGF-β1, α2-macroglobulin, MMP2, ApoA1, urea, TIMP, MMP1 and haptoglobin	Single stage classification model: ANN: tangent sigmoid transfer function. DM: decision tree, chi-square, entropy reduction, gini reduction splitting criteria. Multivariate logistic regression analysis. A multistage stepwise classification model: ANN: linear, tangent sigmoid and logarithmic sigmoid, back-propagation algorithms (gradient descent, gradient descent with momentum,	Prediction of liver fibrosis degree in patients with chronic hepatitis C infection. Accuracy: Single stage model: 82.8% (training), 71.2% (testing) Multistage model: 85.6% (testing), 81.9% (training)

		conjugate-gradient, quasi-newton and levenberge-marquardt), mean square error. DM: decision tree, entropy, gini index, chi-square test support and confidence measures. Multivariate logistic regression analysis, likelihood ratio test, hosmer and lemeshow chisquare goodness of fit tests, variance inflationary factor (VIF) test	

2.3.3 ANN-FL

ANN-FL has emerged as one of the highly used integrated method in liver disease. Literature work shows the integration of ANN-FL as one of the best model which has several benefits such as enhanced accuracy [71], flexibility, improved decision ability [72], robustness [73], [74], simplicity and clarity [75]. This integration makes the system reliable, rapid, more accurate, easy to operate, non-invasive, more economical and more efficient. ANN-FL methodology was used by Comak et al. (2007) where FL is used to pre-process liver disorders dataset and ANN is used to classify [71], by Dogantekin et al. (2009) where fuzzy inference system based ANN is used for classification [73], by Li et al. (2010) where FL is used to compute new attribute values and ANN is used to classify [76], by Ceylan et al. (2011) where FL is used to reduce the number of segments in training pattern and ANN is used for classification [75], by Neshat and Zadeh (2010) where FL is used for clustering and ANN is used for classification [72], by Li and Liu (2010) where FL is applied to calculate the similarity of paired data for every class and attribute; and ANN is used to classify [77], by

Celikyilmaz et al. (2009) where ANN is used to approximate fuzzy classification function parameters of each cluster and FL is used to classify [74], and by Kulluk et al. (2013) where the proposed approach extracts brief and accurate fuzzy classification rules (FCR) from ANNs [78].

Neshat and Zadeh (2010) fuzzy hopfield neural network approach had fast computational power and gained better accuracy than other neural networks like MLP, RBF, GRNN, PNN, LVQ and Hopfield. Li et al. (2010) proved its method superiority to SVM and C4.5 decision tree in terms of classification accuracy. As class imbalance problem in medical datasets diminishes the classification performance of traditional techniques, ANN-FL based approach (Li et al., 2010) balances the data size by over-sampling the minority class and under-sampling the majority class. Comak et al. (2007) proposed medical decision making system attains higher classification accuracy than those of methods mentioned in literature which includes RULES-4, C4.5, naive bayes, BNND, BNNF, SVM with GP, SSVM, RSVM, MLP, PNN, GRNN, RBF, and AIRS. Dogantekin et al. (2009) proposed automatic diagnosis system has better classification performance than RBF, FS-AIRS with fuzzy, FSM with rotations, FSM without rotations, MLP with BP, quadratic discriminant analysis, Weighted9NN, 18NN, ALI, MLP+BP, LDA, MLP, RBF, 1NN, naive bayes and semi-NB, fisher discriminant analysis, LVQ, GRNN, ASR, IncNet, CART, PCA-AIRS and LFC. Li and Liu (2010) proposed kernel attains improved classification accuracy than polynomial and gaussian kernels. Kulluk et al. (2013) proposed fuzzy DIFACONN-miner algorithm yields enhanced results than other fuzzy rule based classification algorithms, namely, $2SLAVE_{sum}$, $FRBCS\_GP_{sum}$, and $GP\text{-}COACH_{sum}$. It also minimizes a few complexity problems. The survey on applicability of ANN-FL for liver disorders is listed in Table 2.7.

Table 2.7: Details of ANN-FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Comak et al. (2007) [71]	Liver disorders dataset	ANN: least square support vector machine, set of linear equations for training. FL: fuzzy weighting	Diagnosing liver disorders. Accuracy: 94.29% Sensitivity: 95% Specificity: 93.33%

		pre-processing, triangular (input and output) membership functions, fuzzy IF-Then rules	
Celikyilmaz et al. (2009) [74]	Liver disorders dataset	ANN: SVM, platt's probability method. FL: classical fuzzy c-means (FCM) clustering. Semi-non-parametric inference mechanism, posterior probabilities from logistic regression, three-way data split cross validation method	Liver disorders diagnosis. Accuracy: 77%
Dogantekin et al. (2009) [73]	Hepatitis dataset	ANN: hybrid learning algorithm (back-propagation for non linear parameters and least square errors for linear parameters) FL: fuzzy IF-Then rules, bell-shaped membership function.	Diagnosis of hepatitis. Accuracy: 94.16% Sensitivity: 96.66% Specificity: 91.66%
Neshat and Zadeh (2010) [72]	Liver disorders dataset	ANN: hopfield neural network, discrete and continuous models,	Liver disorders diagnosis. Accuracy: FHNN-92% HNN-88.2%

		sigmoid activating function. FL: fuzzy c-means	
Li et al. (2010) [76]	Liver disorders dataset	ANN: support vector machine classifier. FL: gaussian type fuzzy membership function and α-cut to reduce data size, mega-trend diffusion membership function, Tenfold cross validation	Deal with class imbalance problem in medical datasets and to enhance classification accuracy in BUPA liver disorders dataset. Accuracy: 86.36%
Li and Liu (2010) [77]	Liver disorders dataset	ANN: support vector machine, class probability based kernel, kernel based on gaussian membership function, decomposition principle, diffusion function technique, mega-trend diffusion technique. FL: triangular type membership function	Classification performance on liver disorders dataset. Accuracy: 70.78%
Ceylan et al. (2011) [75]	Doppler signals of 90 subjects (each subject includes 40 samples)	ANN: complex-valued artificial neural network (CVNN), complex back-propagation (CBP) algorithm,	Liver disorders classification (identify liver as healthy or cirrhosis). Accuracy: 100% Sensitivity: 100%

		complex-valued activation function, real and imaginary components. FL: fuzzy clustering, calculation of FFT (Fast Fourier Transform) values, FCM clustering	Specificity: 100%
Kulluk et al. (2013) [78]	Liver disorders dataset	ANN: feed-forward and recurrent ANNs, trained using differential evolution algorithm. FL: triangular membership function, generates fuzzy rules by touring ant colony optimization (TACO) algorithm, fixed length binary encoding scheme to represent rules. Tenfold cross validation, fitness evaluation by minimum deviation method (MDM)	Classify liver disorders. Accuracy: 85.60%

2.3.4 AIS-FL

AIS-FL methodology was used by Polat et al. (2007) where FL is used for resource allocation and AIS is used for classification [79], by Mezyk and Unold (2011) where AIS is

used for induction of fuzzy rules [80]. Polat et al. (2007) proposed fuzzy-artificial immune recognition system obtains highest classification accuracy among classifiers such as RULES-4, C4.5, Naïve Bayes, BNND, BNNF, SVM, SSVM, RSVM, MLP, PNN, GRNN, AIRS and RBF. It has taken less time in computation, effectively solves problems having large dimensioned feature space and too many classes; and required fewer resources than traditional AIRS which make it more beneficial. Mezyk and Unold (2011) proved their IFRAIS method superior to classifiers such as C4.5, Naive Bayes, K*, Meta END, JRip, and Hyper Pipes in terms of classification accuracy. The survey on applicability of AIS-FL for liver disorders is listed in Table 2.8.

Table 2.8: Details of AIS-FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Polat et al. (2007) [79]	Liver disorders dataset	AIS: artificial immune recognition system (AIRS), clonal selection, affinity maturation, memory cell formation. FL: fuzzy resource allocation, IF-Then rules. Tenfold cross validation method	Liver disorders classification. Accuracy: 83.36%
Mezyk and Unold (2011) [80]	Hepatitis dataset and Liver disorders dataset	Induction of fuzzy rules with an artificial immune system (IFRAIS): sequential covering algorithm and clonal selection algorithm, fuzzy partition inferring based on clonal selection	To assess prediction accuracy of liver disorders in patients. Accuracy: Hepatitis: 93.87% BUPA dataset: 72.34%

		algorithm, only continuous attributes were fuzzified, IF-Then fuzzy rules, paired t-test	

2.3.5 ANN-GA

GA is used with ANNs to enhance classification performance of medical systems [81]. GA eliminates irrelevant and noisy features which decreases the size of network. Integration of ANN-GA overcomes the non-linearity problems and solves the complexity problems of each other. ANN-GA based framework has low computational complexity and simplicity of architecture [82]. Gorunescu et al. (2012) has replaced back propagation algorithm with GA based learning to optimize MLP's weights [81]. ANN-GA methodology was used by Dehuri and Cho (2010) where GA is used to select pertinent features and ANN is used to classify [82], and by Gorunescu et al. (2012) where GA is used to optimize the ANNs synaptic weights and ANN is used to classify. Gorunescu et al.(2012) proposed intelligent system attains improved accuracy, for both complete dataset and reduced dataset, than other machine learning techniques accounted in literature including LN, PNN, RBF, 3-MLP and 4-MLP. This intelligent system is even faster and more effective than 3-MLP and 4-MLP. Results (Dehuri and Cho, 2010) demonstrated that proposed method named as HFLNN outperforms other competing classification methods such as radial basis function network (RBFN) and functional link neural network (FLNN) with back propagation learning. The survey on applicability of ANN-GA for liver disorders is listed in Table 2.9.

Table 2.9: Details of ANN-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Dehuri and Cho (2010) [82]	Liver disorders dataset	ANN: back propagation learning and genetic optimization, trigonometric function.	Diagnosis of liver disorders. Accuracy: 77.6820%

		GA: single point crossover operator, mutation operator, selection. Two fold cross validation, parametric t-test, non-parametric wilcoxon signed rank test	
Gorunescu et al. (2012) [81]	Complete dataset: stiffness, sex, body mass index, glycemia, triglycerides, cholesterol, HLD cholesterol, aspartate aminotransferase, alanin aminotransferase, gama glutamyl transpeptidase, total bilirubin alkaline, phosphatase, prothrombin index, quiq time, prothrombin time ratio, prolonged activ. partial thromboplastin time, haematids, hemoglobin, hematocrit, medium erytrocity volume, avg. erytrocitary hemoglobin, avg. concentration of	ANN: multi-layer perceptron architecture. GA: crossover and mutation. Tandem feature selection mechanism, discriminant function analysis, multiple (linear) regression model (both forward stepwise and backward stepwise), tenfold cross validation, binary tournament selection	Classify liver fibrosis stadialization in chronic hepatitis C. Accuracy: 61.16% (complete dataset) 65.21% (reduced dataset)

	hemoglobin in a red blood cell, thrombocytes, sideraemia, interquartile range. Reduced dataset: stiffness, aspartate aminotransferase, prothrombin index, thrombocytes, sideraemia, interquartile range		

2.3.6 AIS-GA

AIS-GA methodology was used by Ozsen and Gunes (2009) where GA is used to determine weights of attributes that gives minimum classification error and then these weights are used in their own previously developed AIS based system [63]. The classification accuracy of GA-AWAIS based system was superior to both AWAIS and other traditional classifiers mentioned in literature such as Fuzzy-AIRS, AIRS, RSVM, MLP, SSVM, SVM with GP, C4.5, GRNN, Naive Bayes, BNNF, BNND, RBF, RULES-4 and PNN. The survey on applicability of AIS-GA for liver disorders is listed in Table 2.10.

Table 2.10: Details of AIS-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Ozsen and Gunes (2009) [63]	Liver disorders dataset	AIS: attribute weighted artificial immune system (AWAIS). GA: single point crossover, hypermutation, selection.	Classification performance on liver disorders dataset. Accuracy: 85.21%

		Tenfold cross validation method	

2.3.7 ANN-PSO

ANN-PSO methodology was used by Qasem and Shamsuddin (2011) where TVMOPSO based RBF networks are developed [83]. TVMOPSO extended the algorithm to handle multi-objective optimization problems. TVMOPSO was simple, robust, easy to use and easy to implement. Classification accuracy of proposed adaptive evolutionary RBF network algorithm is superior to HMOEN_L2, HMOEN_HN and inferior to RBF network based on MOPSO and NSGA-II algorithms. Advantages of this integration method were stability, consistency, simplicity, enhanced accuracy, better convergence and small standard deviations. The survey on applicability of ANN-PSO for liver disorders is listed in Table 2.11.

Table 2.11: Details of ANN-PSO based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Qasem and Shamsuddin (2011) [83]	Hepatitis dataset	ANN: RBF network, centers of RBF network in hidden layer are initialized from k-means clustering algorithm, weights of RBF network are initialized from the least mean squared algorithm, crowding distance operator. PSO: time variant multi-objective particle swarm optimization (TVMOPSO)	Diagnosis of hepatitis diseases. Accuracy: 82.26% Sensitivity: 88.47% Specificity: 41.92% AUC: 0.652

2.3.8 CBR-DM

CBR-DM methodology was used by Lin (2009) where DM is used to find existence of liver disorders and CBR is used to identify types of liver disorder [84]. CBR participated in problem solving by reducing diagnostic errors and meliorating quality of treatment. Though both CBR and DM could be used in first phase for identifying presence of liver disorder but Lin (2009) has chosen DM technique as it obtained better results in terms of accuracy, sensitivity and specificity. Then in second phase, CBR performed reasonably well in identifying the types of liver disorder. The survey on applicability of CBR-DM for liver disorders is listed in Table 2.12.

Table 2.12: Details of CBR-DM based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Lin (2009) [84]	Age, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, total bilirubin, direct bilirubin, total protein, albumin, gamma-glutamyl transpeptidase, alpha-fetoprotein, sex, blood type, HBsAg, HBeAg, Anti-HBs, Anti-HBe, Anti-HBc, Anti-HCV	CBR: retrieve most similar cases, reuse cases, revise potential solution, retain new solution, assign indices and weights, case adaptation (through inclusion, removal, substitution or transformation). DM: CART, tree-building method, binary tree structure, recursive binary splitting, tree growing and tree pruning stages, gini diversity index Fivefold cross	Phase 1: Liver disorders diagnosis Phase II: Identify types of liver disorders (chronic hepatitis, alcohol hepatitis, liver cirrhosis and B hepatitis. Accuracy: Phase I (CART): Accuracy: 92.94% Sensitivity: 96.00% Specificity: 88.57% AUC: 0.928 Phase II (CBR): Accuracy: 90.00% Sensitivity: 91.09% Specificity: 88.41% AUC: 0.889

		validation method	

2.3.9 CBR-GA

The CBR-GA methodology was used by Park et al. (2011) where CBR uses GA to find optimal cut-off distance and cut-off classification point [85]. This integration overcame the limitation of conventional CBR of being deficient in reflecting asymmetric misclassification cost. It was found that average total misclassification cost of proposed method (Park et al., 2011) is considerably less than C5.0 and CART cost-sensitive learning methods for a number of datasets. The survey on applicability of CBR-GA for liver disorders is listed in Table 2.13.

Table 2.13: Details of CBR-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Park et al. (2011) [85]	Hepatitis dataset	CBR: cost-sensitive case-based reasoning (CSCBR), case retrieval. GA: reproduction, crossover and mutation operators. Tenfold cross validation, paired t-test	Total misclassification cost of CSCBR in medical datasets (Hepatitis). Total Cost: 7.0600

2.3.10 DM-GA

DM-GA methodology was used by Sarkar et al. (2012) where DM is used for producing rules from training dataset and GA is used for handling interpretability problem [86], and by Stoean et al. (2011) where DM is used to extract features and GA is used to build rules for establishing the diagnosis [87]. In comparison, proposed learning system called DTGA (decision tree and genetic algorithm) [86] is more accurate than classifiers such as neural network, naive bayes, C4.5, rough-set based rule inducer; and is less sensitive to missing data compared to NN and C4.5. This integration system also enhanced the performance over volumetric data and had less time complexity compared to majority of GA based approaches.

Cooperative coevolutionary algorithm (CCEA) [87] based proposed technique attains smallest standard deviation and highest accuracy among classification techniques like SVM, NN, SVM + PCA and NN + PCA. The survey on applicability of DM-GA for liver disorders is listed in Table 2.14.

Table 2.14: Details of DM-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Stoean et al. (2011) [87]	Stiffness, sex, body mass index, glycemia, triglycerides, cholesterol, HDL cholesterol, aspartate aminotransferase, alanin aminotransferase, gama glutamyltranspeptidase, total bilirubin, alkaline phosphatase, prothrombin index, INR (prothrombin time ratio), prolonged activated partial thromboplastin time. haematids (erythrocytes), hemoglobin, hematocrit, medium erytrocity volume, avg. erytrocitary hemoglobin, Avg. concentration of hemoglobin in a red blood cell, thrombocytes,	DM: PCA for feature extraction, IF- Then rules. GA: mutation operator, cooperative coevolutionary algorithm (CCEA). Hill climbing algorithm	Liver fibrosis diagnosis (differentiate between five degrees of liver fibrosis). Accuracy: 62.11%

	sideraemia		
Sarkar et al. (2012) [86]	Liver disorders dataset	DM: C4.5 decision tree based rule inducer algorithm, IF-Then rules. GA: selection, single point crossover, mutation	To improve prediction accuracy for liver disorders irrespective to domain and size. Accuracy: 80.02%

2.3.11 DM-FL

The DM-FL methodology was used by Luukka and Leppalampi (2006) where DM is used for dimension reduction and fuzzy similarity model is used for classification [88], by Luukka (2009) where fuzzy robust PCA algorithm is used for data preprocessing and similarity classifier for classification [89], and by Torun and Tohumoglu (2011) where DM is used to group the data and FL is used for classification [90]. The DM-FL integration based systems are robust and effective in diagnosis. These systems also provide semantic information about classification task [88] and had obtained improved accuracies [89]. Luukka and Leppalampi (2006) fuzzy similarity model performed fairly well compared to other classifiers like C4.5-1 (C4.5 with default learning parameters) and C4.5-3 (C4.5 with parameter c equal to 95). Classification accuracies were obtained using both dimension reduction methods: PCA and entropy minimization. The mean classification result was a bit higher using PCA than using entropy. Luukka (2009) proposed FRPCA classification method shows higher accuracy when compared with those of conventional PCA and similarity classifier. Torun and Tohumoglu (2011) proposed simulated annealing and subtractive clustering based fuzzy classifier (SASCFC) in four different versions (SASCFC Type1, Type2, Type3, and Type4). Classifications results obtained from different versions are compared among each other and it was found that SASCFC-Type4 is superior. The survey on applicability of DM-FL for liver disorders is listed in Table 2.15.

Table 2.15: Details of DM-FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Luukka and	Liver disorders dataset	DM: PCA and	Detection of liver

Leppalampi (2006) [88]		entropy minimization method. FL: fuzzy similarity model, lukasiewicz structure for defining memberships of objects	disorders. Accuracy: 66.50% (PCA) 66.06% (entropy)
Luukka (2009) [89]	Liver disorders dataset	Data was preprocessed using Fuzzy robust PCA algorithms (FRPCA), similarity classifier for classification	Classification accuracy on liver disorders data. Accuracy: 70.25%
Torun and Tohumoglu (2011) [90]	Liver disorders dataset	FL: fuzzy IF-Then rules, fuzzy inference system (FIS). DM: subtractive clustering, wrapper type feature selection approach. Simulated annealing, least square estimation, k-fold cross validation	Liver disorders classification. Accuracy: SASCFC- Type1:73.6% SASCFC-Type2:73.9% SASCFC- Type3:73.93% SASCFC- Type4:74.13%

2.3.12 FL-GA

FL-GA methodology was used by Wang et al. (1998) where GA is used for generating optimal set of fuzzy rules and membership functions [91]; and by Chowdhury et al. (2007) where GA is used to simultaneously integrate multiple fuzzy rule sets and their membership function sets [92]. The proposed genetic algorithm based fuzzy-knowledge integration frameworks needed no human intervention during integration process and was scalable.

Accuracy of the framework increased with increase in data size [92]. The shortcomings of the framework were limited precision, and several unresolved issues in the field of knowledge verification [91]. In terms of classification accuracy, proposed genetic algorithm based fuzzy-knowledge integration framework [91] triumphs over other learning methods mentioned in literature such as CN2, C4, IR*, Bayes, Assistant-86, and Diaconis & Efron's. The survey on applicability of FL- GA for liver disorders is listed in Table 2.16.

Table 2.16: Details of FL-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Wang et al. (1998) [91]	Hepatitis dataset	FL: fuzzy knowledge encoding, fuzzy knowledge integration, IF-Then rules, isosceles-triangle functions, parodi and bonelli parameters to represent each membership function. GA: two-substring crossover operator, two-part mutation operator, pittsburgh approach	Hepatitis diagnosis. Accuracy: 91.61%
Chowdhury et al. (2007) [92]	Age, bilirubin, alk phosphate, SGOT, albumin and protime	FL: fuzzy knowledge encoding and fuzzy knowledge integration, fuzzy rules, isosceles-triangle functions, parodi and bonelli parameters to	Hepatitis diagnosis. Accuracy: 96.33%

		represent each membership function. GA: SBMAC (sub population based max-mean arithmetical crossover), dynamic time-variant mutation (TVM), insertion mutation, deletion mutation, novel evolutionary strategy algorithm	

2.3.13 AIS-ANN-FL

AIS-ANN-FL methodology was used by Kahramanli and Allahverdi (2009) where AIS algorithm is deployed to extract rules from hybrid neural network [93]. Generated rules were very accurate but were large in numbers. Classification accuracy of the proposed integration approach is superior to other classification techniques mentioned in literatures such as C-MLP2LN, FSM and CART. The survey on applicability of AIS-ANN-FL for liver disorders is listed in Table 2.17.

Table 2.17: Details of AIS-ANN-FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Kahramanli and Allahverdi (2009) [93]	Hepatitis dataset	Hybrid neural network: artificial neural network, fuzzy neural network, trained with backpropagation algorithm,	Classification on hepatitis dataset. Accuracy: 96.78% Sensitivity: 97.56% Specificity: 93.75%

		fuzzification, defuzzification, weight update method of backpropagation algorithm. Artificial immune systems (AIS) algorithm: extracting rules from hybrid neural network, IF-Then rules	

2.3.14 ANN-CBR-RBR

ANN-CBR-RBR methodology was used by Obot and Uzoka (2009) where case-based technique outputs constitute an input to ANN and results obtained from ANN are assisted to form rule base [94]. Finally, a hybrid inference engine has been built to obtain accuracy through rule base. This hybrid system provides high diagnostic accuracy, and high speed for retrieval of information. Limitations of the system are: restricted explanation capability, ineffective in managing noisy data due to fragility of rules, and it would not like the insertion of new knowledge. The survey on applicability of ANN-CBR-RBR for liver disorders is listed in Table 2.18.

Table 2.18: Details of ANN-CBR-RBR based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Obot and Uzoka (2009) [94]	Nausea, vomiting, fever, body weakness, loss of appetite, diarrhea, itching, convulsion, stupor, headache, tremors, skin discoloration, eye	ANN: trained with multilayer perceptron backpropagation neural networks (MLPBPNN). CBR: retrieval using binary search	Diagnosis of hepatitis disease

	discoloration, liver tenderness, bile in urine, jaundice	algorithm, adaptation. RBR: IF-Then rules	

2.3.15 AIS-DM-FL

AIS-DM-FL methodology was used by Polat and Gunes (2007a) where DM is used for dimensionality reduction, FL is used for data-weighted processing and AIS for classification [95], and by Polat and Gunes (2007b) where DM is used for feature reduction, FL is used for weighting the whole dataset and AIS for classification [96]. Advantage of deploying AIRS was that it is not necessary to know the appropriate settings for the classifier. In terms of classification accuracy, Polat and Gunes (2007a) proposed hybrid approach triumphs over other learning methods such as MLP, RBF (ToolDiag), MLP + BP (ToolDiag) and GRNN. Another Polat and Gunes (2007b) proposed machine learning approach obtains very promising results which are effective, accurate and superior to weighted 9-NN, 18-NN, 15-NN, FSM with rotations, FSM without rotations, RBF (ToolDiag), LDA, naive bayes, QDA, 1-NN, ASR, fisher discriminant analysis, LVQ, CART (decision tree), MLP with BP, ASI, LFC, IncNet, MLP, RBF and GRNN. The survey on applicability of AIS-DM-FL for liver disorders is listed in Table 2.19.

Table 2.19: Details of AIS-DM-FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Polat and Gunes (2007a) [95]	Hepatitis dataset	DM: C4.5 decision tree algorithm, wrapper approach, filter-based feature selection. FL: fuzzy weighted pre-processing, triangular membership functions, fuzzy IF-Then rules.	Diagnosis of hepatitis disease. Accuracy: 81.82% Sensitivity: 55.56% Specificity: 83.82%

		AIS: AIRS supervised learning algorithm, immune mechanisms used are resource competition, clonal selection, affinity maturation and memory cell formation, stages of AIRS includes initialization, memory cell identification and ARB generation	
Polat and Gunes (2007b) [96]	Hepatitis dataset	DM: C4.5 decision tree algorithm. FL: weighted with fuzzy weighted pre-processing, triangular Membership functions, fuzzy IF-Then rules. AIS: AIRS, immune metaphors used are antibody- antigen binding, affinity maturation, clonal selection process, resource distribution and memory acquisition, learning algorithm consists of	Diagnosis of hepatitis disease. Accuracy: 94.12% Sensitivity: 100% Specificity: 92.85% ADR: 96.42%

		initialization, memory cell recognition, resource competition and revision of resulted memory cells. Tenfold cross validation method	

2.3.16 ANN-DM-FL

ANN-DM-FL methodology was used by Su et al. (2006) where fuzzy ART neural network is employed to construct information granules and DM is used to extract knowledge rules from the granules [97], and by Li et al. (2011) where fuzzy-based non-linear transformation method is applied to extend classification related information, DM is used for extracting the optimal subset of features and ANN is used for classification [98]. The hybrid model [97] effectively dealt with imbalanced datasets. Obtained simulated results proves superiority of proposed fuzzy-based non-linear transformation method [98] over PCA and kernel principal component analysis (KPCA); and superiority of proposed knowledge acquisition via information granulation model [97] over C4.5 and SVM. The survey on applicability of ANN-DM-FL for liver disorders is listed in Table 2.20.

Table 2.20: Details of ANN-DM-FL based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Su et al. (2006) [97]	Liver disorders dataset	ANN: Fuzzy ART (Adaptive Resonance Theory) neural network. DM: knowledge acquisition via C4.5 decision tree	Improve classification performance by solving class imbalance problems. Accuracy (KAIG): 70% Data type: Information granules
Li et al. (2011) [98]	Liver disorders dataset	FL: fuzzy-based non-linear transformation	To increase classification performance with small

		method, triangle shape membership function, a fuzzy membership computational approach. DM: principal component analysis. ANN: SVM, megatrend diffusion (MTD) function, polynomial and gaussian kernel. T-test, friedman test, ANOVA test	medical datasets. Accuracy: SVM (poly): 54.21% SVM (gaus): 54.13%

2.3.17 ANN-DM-GA

ANN-DM-GA methodology was used by Stoean et al. (2015) where GA is used to dynamically concentrate search only on most relevant attributes, DM is used to reduce the data dimensionality, and ANN makes the novel model flexible and good performer [99]. Proposed evolutionary approach also obtains better classification accuracy than traditional support vector machines. ESVM was deployed instead of SVM because it has successfully resolved complexity of SVM. ESVM had more simplicity, stability, flexibility, robustness, transparency, adaptability and operability but ESVM training was a bit slow than standard SVMs. The survey on applicability of ANN-DM-GA for liver disorders is listed in Table 2.21.

Table 2.21: Details of ANN-DM-GA based systems with their results and applications

Author , Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Stoean et al. (2015) [99]	Stiffness, sex, BMI (body mass index), glycemia, triglycerides,	ANN: evolutionary-powered support vector machines	Determining the degree of liver fibrosis in chronic hepatitis C.

	cholesterol, HDL cholesterol, aspartate aminotransferase, alanin aminotransferase, gama glutamyl transpeptidase, total bilirubin, alkaline phosphatase, prothrombin index, quiq time, prothrombin time ratio, prolonged activated partial thromboplastin time, haematids, hemoglobin, hematocrit, medium erytrocity volume, avg. erytrocitary hemoglobin, avg. concentration of hemoglobin in a red blood cell, thrombocytes, sideraemia	(ESVM). DM: principal component analysis. GA: tournament selection method, reproduction (randomly selected), recombination (randomly generated), mutation (randomly generated)	Accuracy: 77.31%

2.3.18 ANN-GA-RBR

ANN-GA-RBR methodology was used by Ramirez et al. (2012) where RBR is used for decision making, GA is used for obtaining new offspring and ANN is used for classification [100]. Liver transplantation is the only treatment for patients having incurable liver disorders. Availability of donors is less due to number of requirements; and transplantation is solely dependent on the availability of liver donors. This disproportion may result in countless deaths. Ramirez et al. (2012) did a fruitful research work by developing a donor-recipient decision system for liver transplantation which prioritizes recipients in queue. The tool is intelligible and sensible for physicians. The survey on applicability of ANN-GA-RBR for liver disorders is listed in Table 2.22.

Table 2.22: Details of ANN-GA-RBR based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Ramirez et al. (2012) [100]	A liver transplant dataset composed of 1001 patterns is used for experimentation	ANN: feed-forward neural network, linear basis function, probabilistic function, radial basis function neural network, probability density function of generalized gaussian distribution, trained with a multi-objective evolutionary learning algorithm (MOEA) called MPENSGA2. GA: multi-objective evolutionary algorithms, structural and parametric mutation operators. RBR: rule-based system designed using two ANN models named MPENSGA2-E and MPENSGA2-MS.	Liver transplantation decision

2.3.19 CBR-GA-PSO

CBR-GA-PSO methodology was used by Chang et al. (2012) where CBR is used to preprocess dataset, GA is used to evolve weights of each attribute in PSO and PSO is used to

construct the medical classification system [101]. The proposed framework generates more precise, effective and intelligible results. Advantage of using PSO was its capability to overcome overlapping situation of dataset. Simulated results compared with other forecasting models such as SVM, KNN, Naive Bayes, FDT, RULES-4, C4.5, BNND, BNNF, SVM with GP, SSVM, RSVM, MLP, PNN, and GRNN demonstrate the superiority of proposed model (CBRPSO). It is also proved that this model has the capability to produce high compact clustering than methods like PSO and K-means. Different PSO based approaches are compared in which GA-CBRPSO outperforms PSO and CBR-PSO. The survey on applicability of CBR-GA-PSO for liver disorders is listed in Table 2.23.

Table 2.23: Details of CBR-GA-PSO based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Chang et al. (2012) [101]	Liver disorders dataset	CBR: case base weighted cluster algorithm, weight vector, gradient method, feature evaluation function. GA: selection, crossover, mutation, replacement. PSO: PSO tool evolved by genetic algorithm, global searching stage, local refining stage	Liver disorders diagnosis. Accuracy: 76.8%

2.3.20 DM-FL-GA

DM-FL-GA methodology was used by Leung et al. (2011) where a DM based framework is introduced in which GA is used for searching and optimization and FL is used to increase performance [102]. Proposed data mining framework obtains much better results than other forecasting models mentioned in literature such as SVM, C5.0 decision tree, neural network

and naive bayes. The survey on applicability of DM-FL-GA for liver disorders is listed in Table 2.24.

Table 2.24: Details of DM-FL-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Leung et al. (2011) [102]	Genomic sequences: HBV DNA sequences, either genotype B or C over 200 patients	DM: molecular evolution analysis, phylogenetic tree analysis, information gain criterion, rule learning based on evolutionary algorithm. GA: generic genetic programming (GGP). FL: fuzzy measure for good performance of classification method. Tenfold method	To classify the HBV DNA data into liver cancer (HCC) and normal (CON, control) classes

2.3.21 CBR-DM-FL-GA

CBR-DM-FL-GA methodology was used by Chang et al. (2010) where CBR is used for decomposing the database into a set of smaller database, fuzzy decision tree is used to classify data and GA is used for selecting optimal fuzzy terms which in long term improves accuracy [103], and by Fan et al. (2011) where CBR is used for clustering the dataset, fuzzy decision tree is used to classify data and GA is used for further improving the classified result by evolving the number of fuzzy terms [104]. These integrated systems were more precise and effective; and also productively assisted doctors in diagnosis. It is noticed that CBR-FDT model [103] reaches the highest classification accuracy among other benchmark classifiers such as improved particle swarm optimization model (KNMPSO), SVM, K-nearest neighbor (KNN), SVM, RMSVM and back propagation neural networks (BPN). Average hit rate of

this model was also highest among all mentioned approaches on different database classification applications. Proposed CBFDT model [104] also shows promising performance and obtains higher classification accuracy than RULES-4, C4.5, BNND, BNNF, SVM with GP, SSVM, RSVM, KNN, Naive Bayes, MLP, PNN GRNN, HNFB and SVM. The survey on applicability of CBR-DM-FL-GA for liver disorders is listed in Table 2.25.

Table 2.25: Details of CBR-DM-FL-GA based systems with their results and applications

Author, Year	**Attributes**	**Intelligent techniques and other methods**	**Result and Application**
Chang et al. (2010) [103]	Liver disorders dataset	CBR: case-based weighted cluster algorithm. DM and FL: fuzzy decision tree (FDT) generated from ID 3 algorithm based on recursive binary partitioning algorithm, data fuzzification, triangle membership functions, fuzzy rules. GA: reproduction/selection, replacement. Stepwise regression analysis (SRA) model	Liver disorders diagnosis. Accuracy: 81.6%
Fan et al. (2011) [104]	Liver disorders dataset	CBR: case-based weighted cluster algorithm. DM and FL: fuzzy decision tree	Liver disorders diagnosis. Accuracy: 81.6% (Average) 90.40% (Best)

		(FDT) generated from ID 3 algorithm based on recursive binary partitioning algorithm, triangle membership functions, fuzzy rules GA: representation and selection (tournament method) , two-point crossover method, two-point mutation method, binary code was adopted, de-coding, tournament method, replace, terminate Stepwise regression analysis (SRA) model	

2.4 Observations

Based on all the search results Table 2.26 have been prepared. This table details the applicability of intelligent techniques to different types of liver disease (hepatitis, liver fibrosis, liver cirrhosis, liver cancer, fatty liver, liver disorders dataset, hepatitis dataset, hepatobiliary disorders dataset and others). Liver disorders dataset is used to diagnose liver disorders that might arise from excessive alcohol consumption. Hepatobiliary disorders dataset is used to diagnose alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. Others column includes five articles in total out of which 3 articles are ANN based (Lee et al., 2005; Jeon et al., 2013; Su and Yang, 2008), 1 is DM based (Jen et al., 2012) and 1 is ANN-GA-RBR based (Ramirez et al., 2012). Either of the two conditions are checked before classifying an article in “others” column. First, if it has neither used liver disorders dataset, hepatitis dataset and hepatobiliary disorders dataset nor covered the types of

liver disorder which are considered for this survey. Second, if the work is related to liver other than diagnosis.

Table 2.26 presents that which individual and integrated ITs were appreciably used for what type of liver disease. For hepatitis, ANN, DM, FL, ANN-CBR, CBR-DM, FL-GA and ANN-CBR-RBR were used; for liver fibrosis, ANN, ANN-DM, ANN-GA, DM-GA and ANN-DM-GA were applied; for liver cirrhosis, ANN, DM, FL, EDC, ANN-CBR, ANN-FL and CBR-DM were used; for liver cancer, ANN, FL, EDC and DM-FL-GA were applied; for fatty liver, ANN and FL were used; for liver disorders dataset, ANN, DM, FL, EDC, ANN-DM, ANN-FL, AIS-FL, ANN-GA, AIS-GA, DM-GA, DM-FL, ANN-DM-FL, CBR- GA-PSO and CBR-DM-FL-GA were applied; for hepatitis dataset, ANN, EDC, ANN-DM, ANN-FL, AIS-FL, ANN-PSO, CBR-GA, FL-GA, AIS-ANN-FL and AIS-DM-FL were used; for hepatobiliary disorders dataset, ANN and FL were applied.

Table 2.26: Applicability of intelligent techniques to different types of liver disease

	Liver Disease									Total
	Hepatitis	Fibrosis	Cirrhosis	Cancer	Fatty liver	LDD	HD	HDD	Others	
ANN	3	1	8	4	2	4	5	2	3	32
DM	2		2			1			1	6
FL	1		1	1	1	3		1		8
GA			1	1		2	1			5
ANN-CBR	2		1							3
ANN-DM		1				1	2			4
ANN-FL			1			6	1			8
AIS-FL						2	1			3
ANN-GA		1				1				2
AIS-GA						1				1
ANN-PSO							1			1
CBR-DM	1		1							2
CBR-GA							1			1
DM-GA		1				1				2
DM-FL						3				3
FL-GA	1						1			2
AIS-ANN-FL							1			1
ANN-CBR-RBR	1									1
AIS-DM-FL							2			2
ANN-DM-FL						2				2
ANN-DM-GA		1								1
ANN-GA-RBR									1	1
CBR- GA-PSO						1				1
DM-FL-GA				1						1
CBR-DM-FL-GA						2				2

For comparison among individual ITs, Figure 2.1 presented the number of ANN, DM, FL and EDC based studies. Figure 2.2 illustrated the preference of integrated ITs compared to

each other. It is observed from Figures 2.1 and 2.2 that most of the researchers have preferred to use ANN and ANN-FL methodologies compared to other techniques. It has also found that ANN was mostly integrated with DM and FL, and vice versa. Study shows the negligible applicability of integrated ITs to liver disorders before the year 2007. If we talk in percentage than 80% usage of integrated techniques was after 2007 with just 20% usage till 2006. Another message emerging from the study is that not even a single intelligent technique had been applied between the years 1999 and 2002.

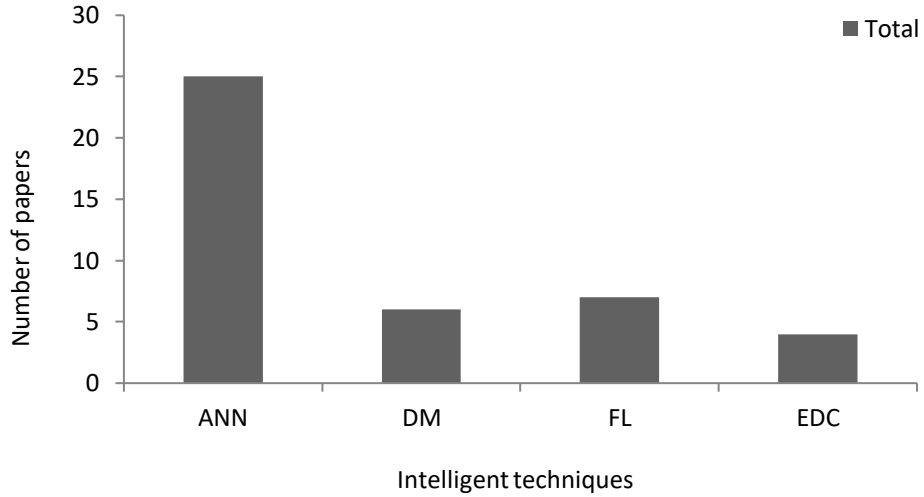

Figure 2.1: Individual intelligent techniques for liver disease

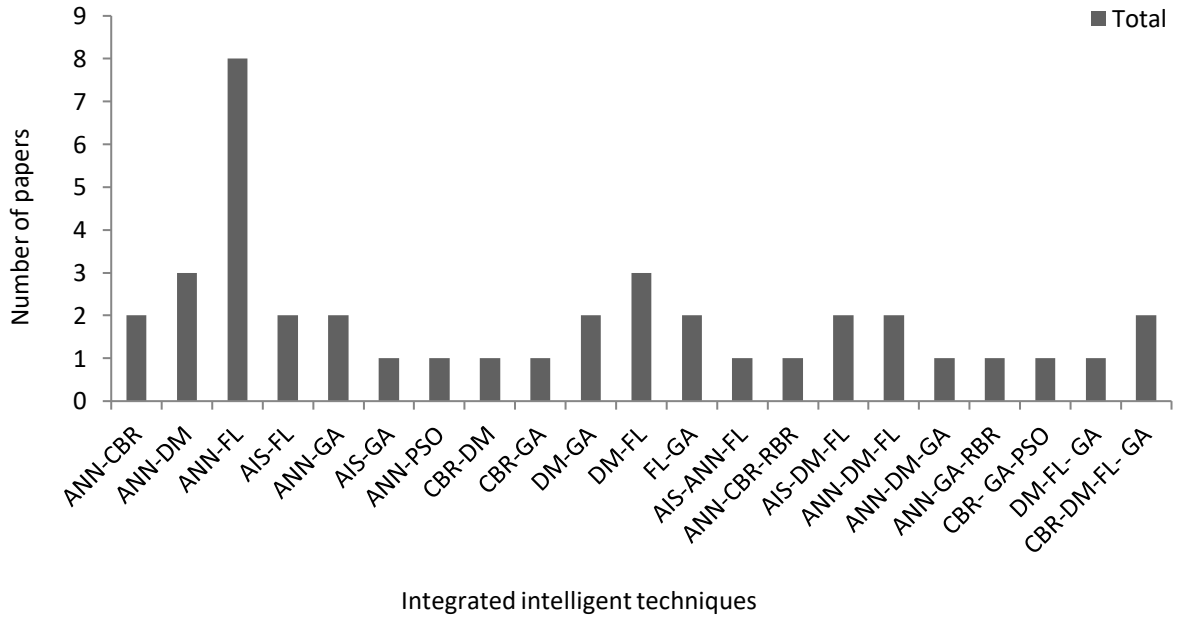

Figure 2.2: Integrated intelligent techniques for liver disease

On computed relative comparison, it was found that ANN method is most significantly used with its 60% rate, FL method is in the second rank with its 17% rate, DM method is in the third rank with its 14% rate, and EDC method is the last one with only 9% rate. In integrated intelligent techniques, mostly used methodology was ANN-FL with 20% rate, then ANN-DM, DM-FL with 8% rate, then AIS-FL, ANN-GA, ANN-CBR, DM-GA, FL-GA, AIS-DM-FL, ANN-DM-FL, CBR-DM-FL-GA with 5% rate and the last ones were AIS-GA, ANN-PSO, CBR-DM, CBR-GA, ANN-CBR-RBR, AIS-ANN-FL, ANN-DM-GA, ANN-GA-RBR, CBR-GA-PSO, DM-FL-GA with 2% rate.

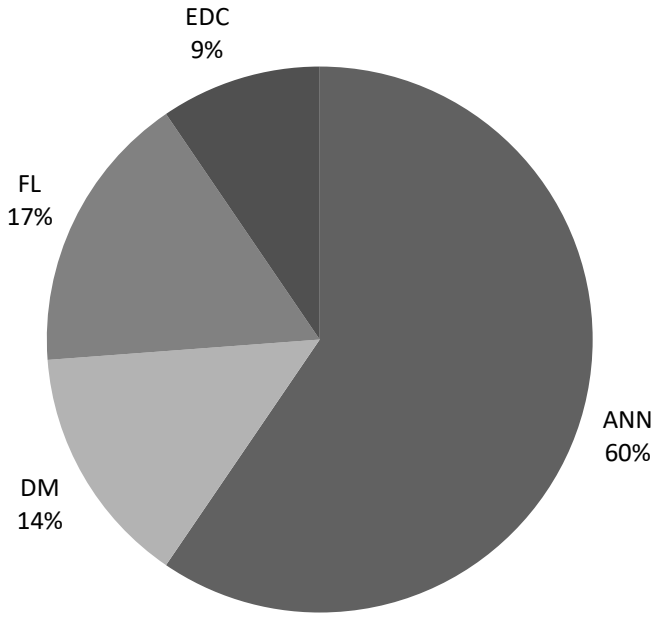

Figure 2.3: Comparative view of percentage use of individual intelligent techniques

During the study we observed that the term intelligent techniques has not been used well enough as keyword in articles. Terms like ANN, AIS, CBR, DM, FL, GA, PSO, RBR, hybrid systems and integrated methods are mostly chosen instead of ITs. So we have introduced a new keyword “Intelligent techniques” which refers to all methodologies mentioned in this study. Along with number of benefits, doing literature review also has several limitations like author’s limited knowledge as one needs to have extensive background information for accumulating, studying and classifying articles; number of papers may have used intelligent techniques but still left out due to some indexing problems; publications using languages other than English cannot be included; it is difficult to cite all academic articles that are listed in science citation index as the amount of available text is increasing rapidly; and we do have limited access to online databases and also bounded by time constraint.

Hopefully, this study would be productive for neophyte researchers, about what to do and what not to, in developing medical decision-making systems for assisting physicians in evaluating liver disease. Novice researchers can either use methodologies like ANN, ANN-FL which has wide acceptance and has obtained higher accuracy results or can choose techniques like AIS, PSO which have not been explored enough hitherto and can help in improving results either alone or when integrated with some other intelligent technique. For hepatitis diagnosis, ANN and ANN-CBR were more preferable; for liver cirrhosis diagnosis, ANN and DM were more suitable; for liver cancer and fatty liver diagnosis, ANN was more used; for liver disorders dataset, ANN-FL was more applied; for hepatitis dataset and hepatobiliary disorders dataset, again ANN was more deployed.

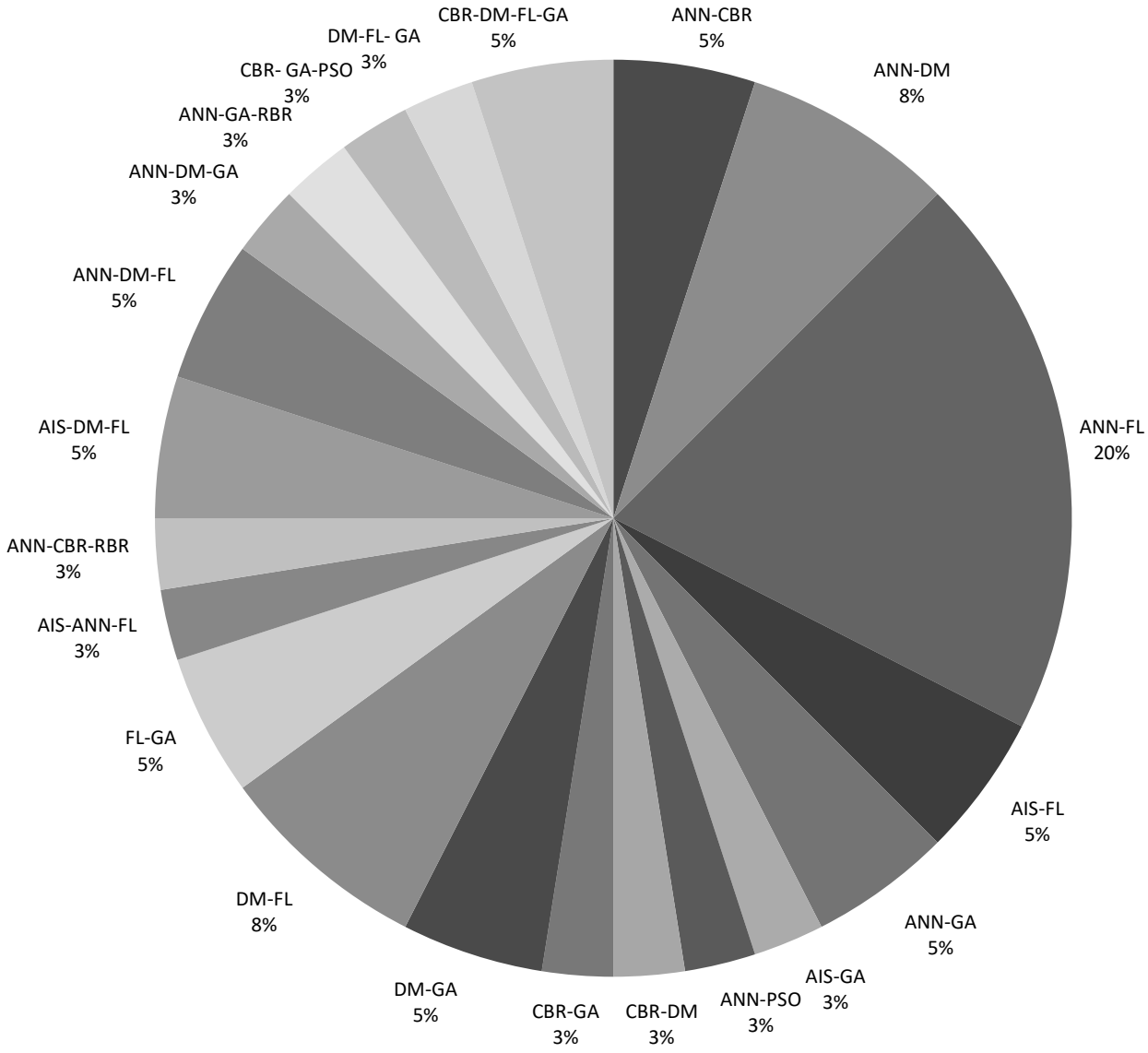

Figure 2.4: Comparative view of percentage use of integrated intelligent techniques

2.5 Conclusions

This chapter has made a contribution to medical field by presenting a study on intelligent techniques applied to liver disease. To the best of our knowledge, not a single attempt had been made to write review on liver disorders for the last 40 years (1976–2016). Numerous authors have written literature review sections but no complete study has been found so far. This study would be helpful for researchers in developing efficient decision-making tools, as one need to be well acquainted with the applicability of ITs to liver disease and also about which method is widely applied for what type of liver disorder. The different types of liver disorder covered in this chapter are: hepatitis, liver fibrosis, liver cancer, fatty liver, general liver damage, alcoholic liver damage, primary hepatoma, cholelithiasis and liver cirrhosis. Articles were searched using different keyword indices such as “liver disorders”, “liver disorders diagnosis”, “hepatitis”, “liver fibrosis”, “liver cirrhosis”, “liver cancer”, “fatty liver”, “ANN used for liver disorders”, “AIS used for liver disorders”, “CBR used for liver disorders”, “DM used for liver disorders”, “FL used for liver disorders”, “GA used for liver disorders”, “PSO used for liver disorders” and “RBR used for liver disorders”; and then classified based on intelligent techniques applied and for what types of liver disorder. Trends indicating from the survey tables were that all intelligent techniques had been progressively applied, from 2007, to liver disease. The study also discovered the merits and demerits (if any) of proposed medical systems that are developed using individual and integrated ITs. Systems developed using ITs efficiently handled imprecise, unstructured and dynamic data of patients, and also assisted physicians by acting as a second opinion tool in decision making process of liver diagnosis.

Optimistically, this study has attained the objective of a review in the following manner: it provides detail about the articles published for liver diseases; it presents the information about which individual and integrated techniques are used for what type of liver disorders; which ITs outperformed others in comparison and what are the attributes taken for experimentation; it specifies the articles published with their results and applications; it portrays accurately the characteristics of ITs and compares their usage among each other; it narrows the researcher’s work as they become aware with pros and cons of the intelligent techniques; it also plays a fundamental role in formulating research hypothesis, preparing research design, and collecting and analyzing the data. It is suggested that novice researchers can use methodologies like ANN, ANN-FL which has wide acceptance and has obtained high

accuracy results or can choose techniques like AIS, PSO which have not been explored enough hitherto and can give advance results either alone or when integrated with some other intelligent techniques. For hepatitis diagnosis, ANN and ANN-CBR were more preferable; for liver cirrhosis diagnosis, ANN and DM were more suitable; for liver cancer and fatty liver diagnosis, ANN was more used; for liver disorders dataset, ANN-FL was more applied; for hepatitis dataset and hepatobiliary disorders dataset, again ANN was more deployed. Though there are number of other integration methodologies like AIS-GA, ANN-PSO, CBR-DM and CBR-GA which have not been widely implemented uptil now. So it is completely a researcher's decision to decide with which techniques to proceed based on his background knowledge and the information he has grabbed from this study. It is hoped and anticipated that the humble effort made in this study will assist in the accomplishment of developing accurate and precise decision making tools to diagnose liver disease.

The detailed literature review presented in this chapter has been published in International Journal of Biomedical Engineering and Technology, Vol. 16, No. 1, pp. 27-70, 2014, Inderscience Publishers, United Kingdom, DOI: 10.1504/IJBET.2014. 065638, mentioned under list of publications at the end of chapter 8.

Chapter 3

Detection of Liver Disease Using a Novel Integrated Method Based on Principal Component Analysis and K-Nearest Neighbor

This chapter is organized as follows: Section 3.1 presents the introduction. Section 3.2 describes the methodology used to diagnose liver disease patients. Section 3.3 details the material used, discusses the experimental results and compare the performance of proposed approach with other classification methods. Finally, conclusions are drawn in Section 3.4.

3.1 Introduction

Talk about organ failure and people immediately recall kidney disease. On contrary, there is no such alertness about liver disease and its failure despite the fact that it is one of the leading cause of mortality around the globe. Liver is one of the most vital part and is the largest internal organ in human body. It carries out several metabolic functions like producing bile, making certain proteins for blood clotting, filtering blood, helping in fat digestion, decomposing red blood cells and most prominently detoxifying harmful chemicals [24]. Liver disease is defined as the improper functioning of complex metabolic functions which further leads to serious health ramifications. Liver disease can be acute (for short time) or chronic (for long time) that might put the life at risk [4]. It is generally caused by accumulation of fat in excess, inherited disorders, virus infected damaged hepatocytes, bacteria or fungi, contaminated food and acute consumption of alcohol or drugs [2], [3], [84]. Its wide and hidden presence worldwide makes it a serious area of concern in medicine. Liver disease has been persistently listed as one of the top ten fatal diseases around the globe costing millions of lives every year. Ability of liver to function normally even when partially damaged resists its early presence and makes it more alarming as by then it has suffered significant or permanent eternal damage. This designates that the early diagnosis of liver disease is crucial so that timely treatment can be initiated [2], [4]. Medical interpretations from clinical and laboratory data is a highly demanding task in liver disease diagnosis. The task becomes even more complex if the existing figures are fuzzy. Analyzing these uncertain medical records of patients' stretches the decision time of physicians even if they are experienced and if they are novice then it may take years for them to gain substantial expertise. Moreover, the accurate

diagnosis is still not guaranteed as humans are prone to errors no matter whatever may the reason be like abundant clinical workload or a poor health.

In recent years, computer-aided medical diagnostic systems have been widely practiced in hospitals and are comprehensively assisting physicians in analyzing patients' therapeutic history. Large data centers are created with the use of hardware and software technologies for resourcefully storing medical records in great amount. For experimentation and learning, intelligent classification techniques are applied on these records which can be quickly retrieved any time with the help of computer processing systems. It is proved from literature study that each classifier has its own significance in providing comprehensive information as per the scalability and diversity of data. Each classifier follows unique steps for data processing and computation which makes them distinct in producing results. Hence, to interpret multifaceted dataset, to avoid clinical inexperience and to reduce the evaluation time; this chapter proposes an efficient diagnosis method for detection of liver disease using an integration of principal component analysis and k-nearest neighbor. The method works with combination of feature extraction and classification performed by PCA and KNN respectively. The chapter also deployed linear discriminant analysis, diagonal linear discriminant analysis, quadratic discriminant analysis, diagonal quadratic discriminant analysis and least squares support vector machine classifiers in integration with principal component analysis. Performance of all intelligent integrated methods is compared in terms of accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Results showed that the PCA-KNN based diagnostic method obtains better prediction outcomes than other integrated approaches. In addition to higher accuracy rates, the method also attained remarkable sensitivity and specificity which were a challenging task given an uneven variance among attribute values in the patients' data.

3.2 Methodology

Figure 3.1 illustrates the block diagram of proposed liver diagnostic method where first step is loading the dataset, second is data preprocessing, third is data partitioning, fourth is performing feature extraction, fifth is classifying the samples, and last step is performance evaluation and comparison which eventually decides the best liver diagnostic method. In data preprocessing, each sample is symbolized as a vector of real numbers before giving input to a classifier. For instance, in selector field a sick class is represented by 1 and a healthy class is indicated with 0. The cross validation method divides the observations into training and

testing data, where M observations are randomly selected as an evaluation set. Feature extraction is carried out using principal component analysis where finally four features are extracted for giving input to the classifiers. Intelligent integrated approaches implemented in this chapter are PCA-LDA, PCA-DLDA, PCA-QDA, PCA-DQDA, PCA-LSSVM and PCA-KNN. The prediction performance is calculated using statistical parameters include accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Description of PCA and all classification algorithms deployed to identify presence of liver disease are as follows.

LDA is originally developed by R. A. Fisher in 1936. It works on the concept of searching for a linear combination of variables that best separates two classes. These variables are the predictors, and the classes are the actual targets in numerical form. It is a classification method based on covariance matrix. It works efficiently for disproportionate within-classes frequencies by maximizes the ratio of between-classes variance to within-classes variance for drawing decision region between the given classes [105]–[107]. For example, let's assume that the dataset have X classes; class j mean vector is μ_j where $j=1, 2, \ldots X$; N_j indicates number of samples within class j where $j=1, 2, .. X$.

$$N = \sum_{j=0}^{X} N_j \tag{3.1}$$

$$M_a = \sum_{j=1}^{X} \sum_{i=1}^{N_i} (y_i - \mu_j)(y_i - \mu_j)^T \tag{3.2}$$

$$M_b = \sum_{j=1}^{X} (\mu_j - \mu)(\mu_j - \mu)^T \tag{3.3}$$

$$\mu = 1/X \sum_{j=1}^{X} \mu_j \tag{3.4}$$

where N is defined as the total number of samples, M_a is the within-class scatter matrix, M_b is the between-class scatter matrix and μ is the mean of entire dataset. On the other hand, DLDA is the extension of linear discriminant analysis where covariance matrices are assumed equal across groups.

QDA is considered as the more generalized version of LDA which is used for heterogeneous variance-covariance matrices. It calculates a quadratic score function for each

of the groups. This function belongs to the mean vectors of population and the variance-covariance matrices for *j*th group. The parameters are estimated by maximizing joint likelihood of features and their classes. On the other hand, DQDA is the extension of quadratic discriminant analysis where covariance matrices are used in which all off-diagonal elements are set to be zero [108], [109].

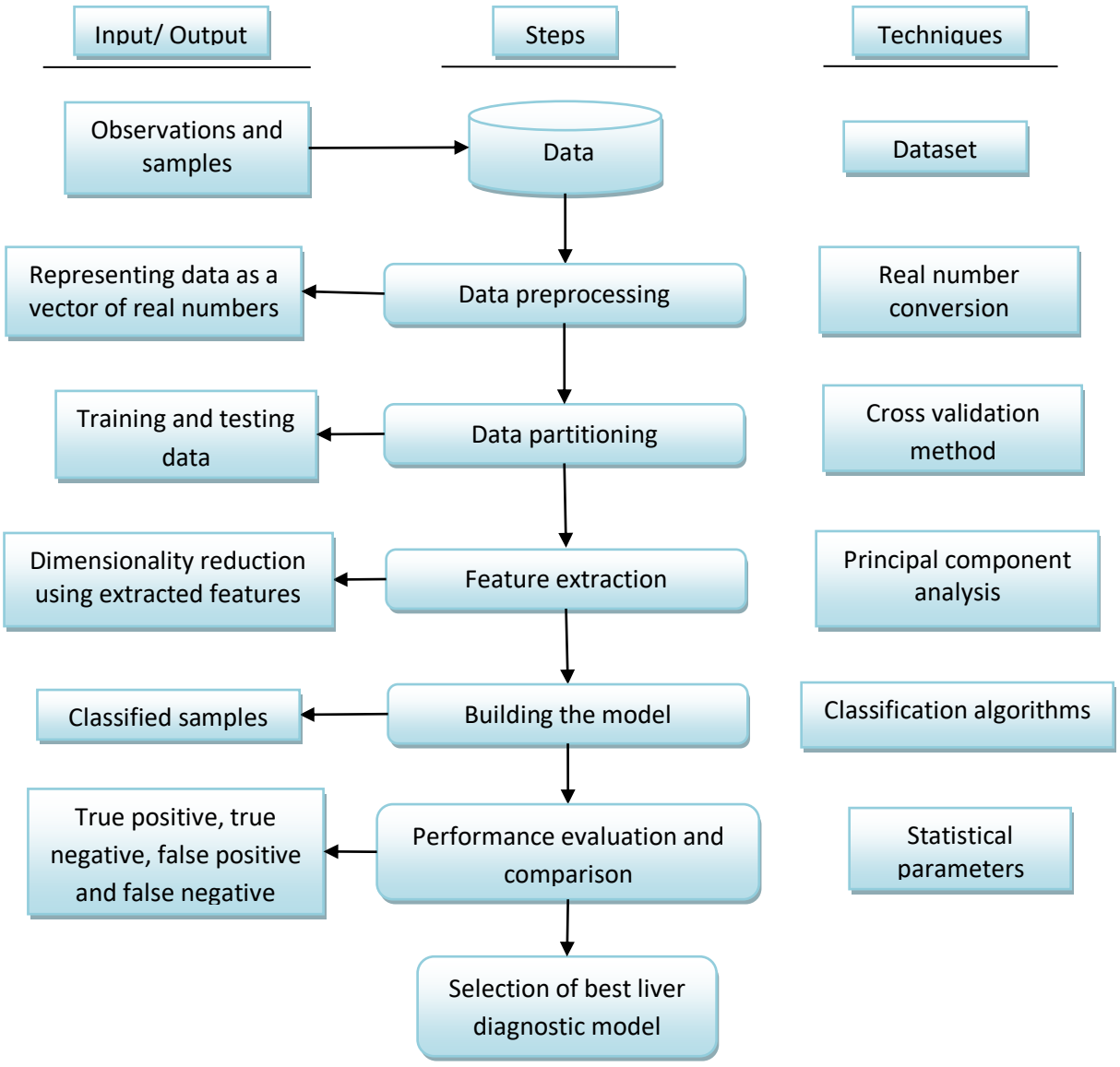

Figure 3.1: Block diagram of proposed diagnostic method for liver disease

Support vector machine based approaches are well known in medical diagnosis and have been extensively used in bioinformatics and pattern recognition. SVMs are initially presented by Vapnik and Cortes for classification and regression tasks [110]. SVM is a supervised learning algorithm that uses kernel functions to fit training data on a suitable hyperplane

surface. Setting the finest kernel parameters leads to constructive performance for the problem. The hyperplane surface is established by support vector machine to separate a range of classes given in the problem. Maximum euclidean distance to the nearest point was used to attain the required hyperplane. Width of gaussian kernels and regularization factor were the two parameters included in LSSVM having a varied value between 1 and 100000 for the former and 0.1 and 10 for the later. Structure of generic SVM shows the distinct classes of data as [+1, -1]. Assume that d data is contained in training data $(p_1, q_1), \ldots., (p_d, q_d),\ p \in R^d\ and\ q \in \{+1, -1\}$. Now, to find optimal hyperplane with maximum possible margin are formed with:

$$H(p) = (z.p) + z_0 \tag{3.5}$$

and the disparity for q = +1 and q = -1 is

$$q_i[(z.p_x) + z_0] \geq 1, x = 1, \ldots., d \tag{3.6}$$

This formula conferred by the data point p and q in equality condition are known as support vectors. The disparity abided by hyperplane margins are as follows.

$$\frac{ql \times H(pl)}{||z||} \geq \Gamma,\ \ l = 1, \ldots., d \tag{3.7}$$

Now, the equation (3.8) is resolute for dropping total number of solutions. The given formula is

$$\Gamma \times \big|\big||z|\big|\big| = 1 \tag{3.8}$$

$$\frac{1}{2}||z||^2 \qquad (3.9),\ \ this\ is\ minimised\ subject\ to\ equation\ (3.6)$$

In case of non-separable data slack variablesβ_i, are added in equation (3.6) and (3.9). The mentioned case then replaces the equation (3.6) and (3.9) with the new one as follows.

$$q_i[(z.p_x) + z_0] \geq 1 - \beta_x \tag{3.10}$$

$$T\sum_{x=1}^{d} \beta_x + \frac{1}{2}\ ||z||^2 \tag{3.11}$$

This generic SVM functioning is productive in the linear classification of data but could not solve the nonlinear cases. To overcome this limitation kernel functions are being built which maps the given data into a kernel space. These kernel functions include linear kernels, quadratic kernels, radial basis function (RBF) kernels, polynomial kernels and multilayer

perceptron (MLP) kernels. The study achieved best results with RBF kernels and least squares method in which the later was used to find the separating hyperplane [111]. RBF kernels is represented as $K(p,p') = \exp(-||p-p'||^2/\sigma^2)$. SVM uses a quadratic optimization problem for training which differs it from LSSVM in which a set of linear equations are used [112]. Like the concept defined in equation (3.10) and (3.11) for SVM, the same can be represented for LSSVM in the following form.

$$For\ x = 1, \ldots., d \qquad q_i[(z.p_x) + z_0] = 1 - \beta_x \tag{3.12}$$

$$\frac{1}{2}\ ||z||^2 + \frac{T}{2}\sum_{x=1}^{d} \beta_x^2 \tag{3.13}$$

$$S(z,c,\alpha,\beta) = \frac{1}{2}\ ||z||^2 + \frac{T}{2}\sum_{x=1}^{d} \beta_x^2\ - \sum_{x=1}^{d} \alpha_x\ \{q_x[(z.p_x) + z_0] - 1 + \beta_x\} \tag{3.14}$$

The α_x is knows as Lagrange multipliers which is an approach to find local maxima and minima for a function subject to a constraint. It should be positive in generic support vector machines structure but can be positive or negative in least squares support vector machine.

KNN is a semi-supervised and competitive learning method that belongs to the family of instance based algorithms. It creates its model based on training dataset and predicts a new data case by searching training data for the k-most similar cases. It strongly retains all observations selected at the time of training. This prediction data case of k-most similar cases is recapitulated and returned as the forecast for a new case. The selection of distance metric functions for finding similarity measure depends on structure of data. Available functions are euclidean, cityblock, cosine, correlation and hamming out of which correlation performed best for this study and hamming was not supportive as it can only be used for categorical or binary data [113]–[115]. Let's assume a pa-by-q data of metric A that can be represented as pa (1-by-q) row vectors a_1, a_2, ... ap_a, and pb-by-q data of metric B that can be represented as pb (1-by-q) row vectors b_1, b_2, ... bp_b. The correlation distance is the statistical difference between vector a_u and b_v are defined in Eq. (3.15).

$$d_{uv} = \left(1 - \frac{(a_u - \bar{a}_u)(b_v - \bar{b}_v)'}{\sqrt{(a_u - \bar{a}_u)(a_u - \bar{a}_u)'}\sqrt{(b_v - \bar{b}_v)(b_v - \bar{b}_v)'}}\right) \tag{3.15}$$

$$where\ \bar{a}_u = \frac{1}{q}\sum\nolimits_j a_{uj} \tag{3.16}$$

$$\bar{b}_v = \frac{1}{q}\sum\nolimits_j a_{vj} \tag{3.17}$$

PCA is an appearance based technique widely used in image recognition, image compression, signal processing, and face recognition. It shows considerable performance in reducing dimensions of a dataset which eventually enhances the results [116]. Therefore, we integrated it with a number of classifiers and finally proposed a PCA-KNN based diagnostic method. PCA working is depicted as mentioned. Assume a dataset D having v dimensions. The p principal axes $Q_1, Q_2, .. Q_p \; where \; 1 \leq p \leq v$. The covariance matrix would be represented as:

$$D = \left(\frac{1}{R}\right) \sum_{i=1}^{R} (c_i - d)^V (c_i - d) \tag{3.18}$$

$where \;\; c_i \in D, d \; is \; samples \; mean, R \; is \; number \; of \; samples.$

$$BQ_i = w_i Q_i \tag{3.19}$$

$where \; i \in 1, \ldots \ldots, p; \; w_i \; is \; ith \; largest \; eigen \; value \; in \; B.$

Finally, p principal components given a $x_i \in D$ can be explained as follows:

$$h = \left[h_1, h_2, \ldots, h_p\right] = [Q_1^V c, Q_2^V c, \ldots, Q_n^V] = Q^V c \tag{3.20}$$

here, h is the p principal component of x.

3.3 Results and Discussion

Liver patient dataset obtained from University of California repository of machine learning databases is used for experimentation. Dataset used in this study has the objective of improving the ability of diagnosing liver disease based on attributes collected. The dataset characteristic is multivariate and it includes 10 attributes, 2 classes and 583 samples. Attributes contain information about age, gender, total bilirubin, direct bilirubin, albumin and globulin ratio, alkaline phosphotase, albumin, alamine aminotransferase, aspartate aminotransferase and total proteins. Data samples with attribute values are given in Table 3.1. Dataset contains two target classes (sick and healthy). Sick class has 416 instances and healthy class has 167. Each line in the data file constitutes record of a single male or female. For record, data has 441 male and 142 female cases. To reduce sample biasness and to estimate misclassification probabilities the patient data was divided into training and testing sets using leave-m-out cross validation method. Obtained results are compared using accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) rates that are defined in Eq. (3.21), (3.22), (3.23), (3.24) and (3.25) respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.21}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.22}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.23}$$

$$PPV = \frac{TP}{TP + FP} \tag{3.24}$$

$$NPV = \frac{TN}{TN + FN} \tag{3.25}$$

where TN indicates true negative (normal people correctly recognized as normal), TP is true positive (diseased people correctly recognized as diseased), FN is false negative (diseased people incorrectly identified as normal), and FP expresses false positive (normal people incorrectly identified as diseased).

Table 3.1: The features of liver patient dataset

Attribute	Description	Values of attribute
Age	Age of the patient	10, 20, 30, 40, 50, 60, 70, 80
Gender	Gender of the patient	Male, female
TB	Total bilirubin	0.7, 10.9, 7.3, 1, 3.9, 6.8
DB	Direct bilirubin	0.1, 1.3, 4.1, 0.4, 2, 0.7, 5.5
ALP	Alkaline phosphatase	490, 187, 195, 208, 154, 699
Sgpt	Alanine aminotransferase	18, 100, 68, 20, 59, 875
Sgot	Aspartate aminotransferase	18, 100, 68, 20, 59, 731
TP	Total proteins	6.8, 7.5, 7, 6.8, 7.3, 7.6
ALB	Albumin	3.3, 3.2, 3.3, 3.4, 2.4, 4.4
A/G ratio	Albumin and globulin ratio	0.9, 0.74, 0.89, 1, 0.4, 1.3
Selector	Field used to split data into two sets	1-sick and 2-healthy

Experimental results showed that PCA-LDA, PCA-DLDA, PCA-QDA, and PCA-DQDA based computational methods had not shown significant diagnostic performance. Although, PCA-LSSVM showed enhanced accuracy rates then aforesaid methods but PCA-KNN achieved highest among all. Figure 3.2, 3.3, 3.4, 3.5 and 3.6 present comparisons among intelligent integrated methods using accuracy, sensitivity, specificity, PPV, and NPV rates respectively. Figure 3.2 illustrates that PCA-LDA had 61.1% (training) and 60.89% (testing) accuracy, PCA-DLDA had 62.13% (training) and 61.92% (testing) accuracy, PCA-QDA had 52.15% (training) and 52.14% (testing) accuracy, PCA-DQDA had 52.67% (training) and 52.66% (testing) accuracy, PCA-LSSVM had 76.08% (training) and 75.99% (testing) accuracy, and PCA-KNN had 100% (training) and 99.83% (testing) accuracy. Figure 3.3

depicts that PCA-LDA had 78.92% (training) and 78.44% (testing) sensitivity, PCA-DLDA had 77.11% (training) and 76..65% (testing) sensitivity, PCA-QDA had 94.58% (training) and 94.61% (testing) sensitivity, PCA-DQDA had 96.39% (training) and 96.41% (testing) sensitivity, PCA-LSSVM had 26.51% (training) and 26.35% (testing) sensitivity, and PCA-KNN had 100% (training) and 100% (testing) sensitivity.

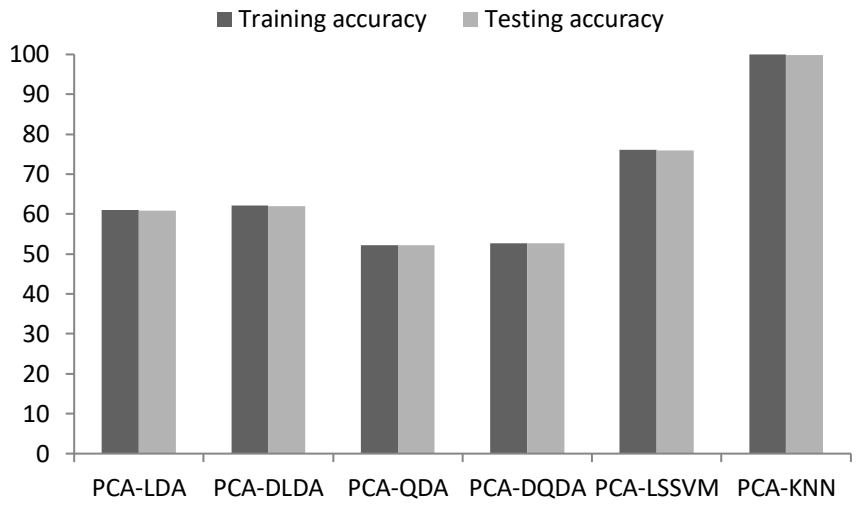

Figure 3.2: The comparative view of obtained accuracy rates

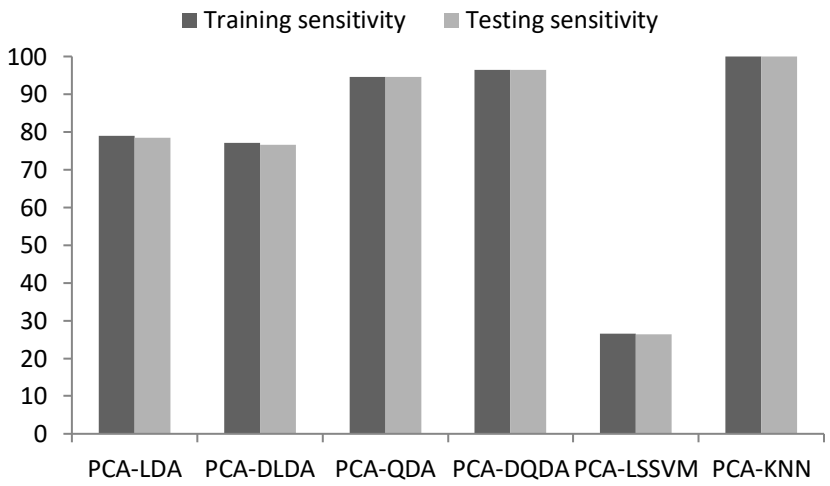

Figure 3.3: The comparative view of obtained sensitivity rates

Figure 3.4 shows that PCA-LDA had 100% (training) and 97.84% (testing) specificity, PCA-DLDA had 56.14% (training) and 56.01% (testing) specificity, PCA-QDA had 35.18% (training) and 35.1% (testing) specificity, PCA-DQDA had 35.18% (training) and 35.1% (testing) specificity, PCA-LSSVM had 95.9% (training) and 95.91% (testing) specificity, and

PCA-KNN had 100% (training) and 99.4% (testing) specificity. Figure 3.5 illustrates that PCA-LDA had 40.68% (training) and 40.56% (testing) PPV, PCA-DLDA had 41.29% (training) and 41.16% (testing) PPV, PCA-QDA had 36.85% (training) and 36.92% (testing) PPV, PCA-DQDA had 37.3% (training) and 37.35% (testing) PPV, PCA-LSSVM had 72.13% (training) and 72.13% (testing) PPV, and PCA-KNN had 100% (training) and 99.76% (testing) PPV. Figure 3.6 demonstrates that PCA-LDA had 86.49% (training) and 86.15% (testing) NPV, PCA-DLDA had 85.98% (training) and 85.66% (testing) NPV, PCA-QDA had 94.19% (training) and 94.19% (testing) NPV, PCA-DQDA had 96.05% (training) and 96.05% (testing) NPV, PCA-LSSVM had 76.54% (training) and 76.447% (testing) NPV, and PCA-KNN had 100% (training) and 100% (testing) NPV.

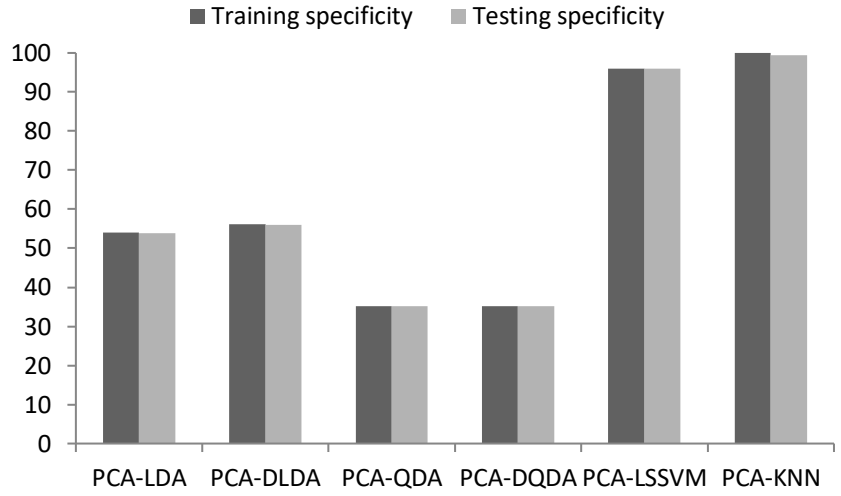

Figure 3.4: The comparative view of obtained specificity rates

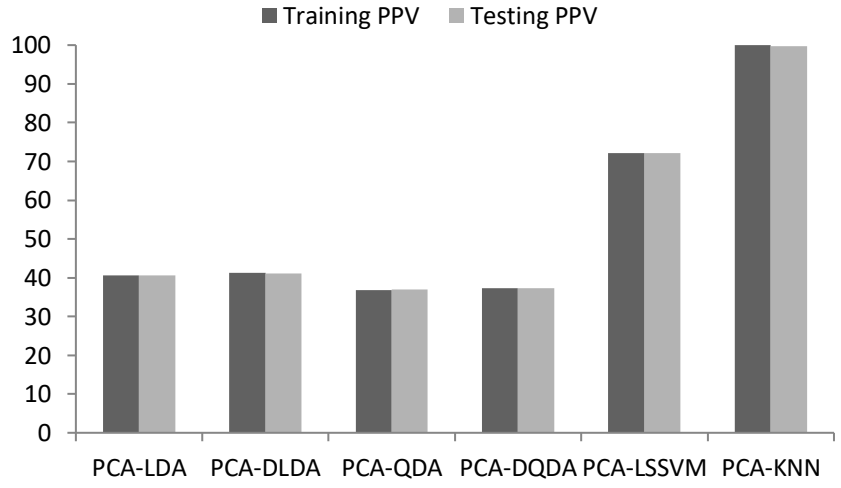

Figure 3.5: The comparative view of obtained positive predictive rates

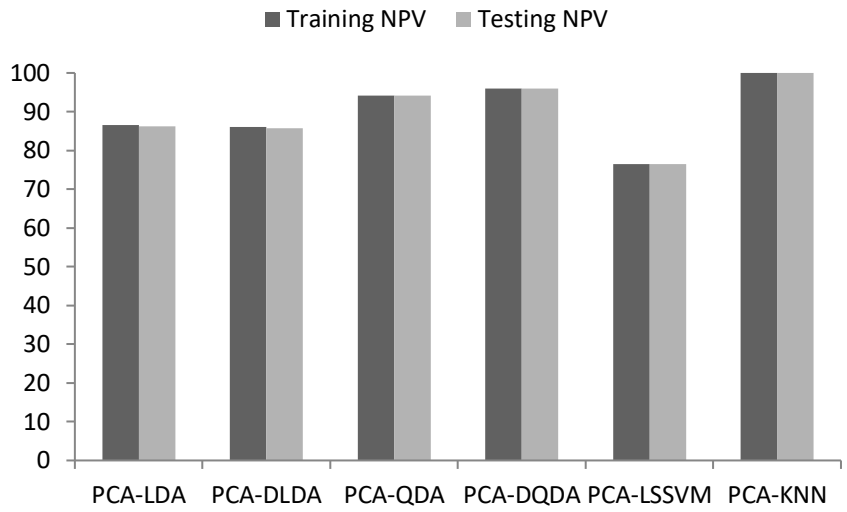

Figure 3.6: The comparative view of obtained negative predictive value rates

SVM is a supervised learning algorithm which uses kernel functions to fit the training data on a suitable hyperplane surface. These kernel functions include linear kernels, quadratic kernels, radial basis function (RBF) kernels, polynomial kernels and multilayer perceptron (MLP) kernels. Setting finest kernel parameter leads to constructive performance for the given problem. Hyperplane surface is established by support vector machine to separate a range of classes given in the problem. Each mentioned kernel function was deployed with least squares (LS), quadratic programming (QP) and sequential minimal optimization (SMO) separating hyperplanes for checking the variations in diagnostic performance of SVM. The study selected RBF kernel function and least squares hyperplane based SVM approach as it outperformed linear, polynomial, quadratic and multilayer perceptron based SVM classifiers.

Figure 3.7 illustrates the comparative view of obtained diagnostic accuracy rates using SVM based approaches. SVM with quadratic kernel function and LS hyperplane had 84.5% (training) and 83.48% (testing) accuracy, SVM with quadratic kernel function and QP hyperplane had 82.9% (training) and 82.13% (testing) accuracy, SVM with quadratic kernel function and SMO hyperplane had 81.81% (training) and 81.22% (testing) accuracy, SVM with linear kernel function and LS hyperplane had 84.09% (training) and 83.95% (testing) accuracy, SVM with linear kernel function and QP hyperplane had 83.79% (training) and 83.67% (testing) accuracy, SVM with linear kernel function and SMO hyperplane had 83.49% (training) and 83.38% (testing) accuracy, SVM with polynomial kernel function and LS hyperplane had 84.01% (training) and 82.04% (testing) accuracy, SVM with polynomial kernel function and QP hyperplane had 82.66% (training) and 81.17% (testing) accuracy,

SVM with polynomial kernel function and SMO hyperplane had 81.79% (training) and 81.71% (testing) accuracy, SVM with MLP kernel function and LS hyperplane had 83.75% (training) and 83.47% (testing) accuracy, SVM with MLP kernel function and QP hyperplane had 82.96% (training) and 82.7% (testing) accuracy, SVM with MLP kernel function and SMO hyperplane had 82.29% (training) and 82.13% (testing) accuracy, SVM with RBF kernel function and LS hyperplane had 86.08% (training) and 85.63% (testing) accuracy, SVM with RBF kernel function and QP hyperplane had 84.85% (training) and 84.69% (testing) accuracy, and SVM with RBF kernel function and SMO hyperplane had 84.6% (training) and 84.46% (testing) accuracy. Finally, LSSVM (SVM with RBF kernel function and LS separating hyperplane) was found superior to other SVM based approaches for liver disease prediction.

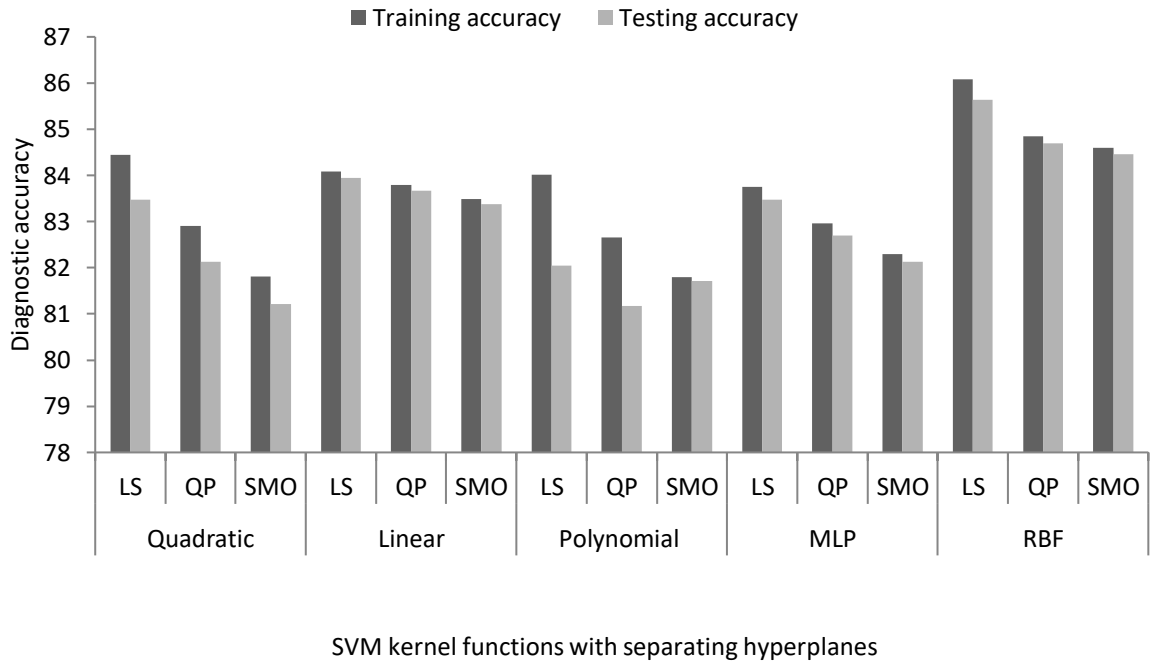

Figure 3.7: The comparative view of diagnostic accuracy rates of SVM based approaches

KNN is a semi-supervised and competitive learning method which belongs to the family of instance based algorithms. It creates its model based on training dataset and predicts a new case by searching training data for the k-most similar cases. It strongly retains all observations selected at the time of training. This prediction data case of k-most similar cases is recapitulated and returned as the forecast for a new case. The selection of distance metric function for finding similarity measure depends on structure of data. These functions include euclidean, cityblock, cosine, correlation and hamming out of which correlation performed best on the given dataset [30, 31]. Each mentioned distance function was deployed with

nearest, random and consensus rules for checking variations in diagnostic performance of k-NN. Based on the obtained results, correlation distance metric function and nearest rule based KNN approach was selected.

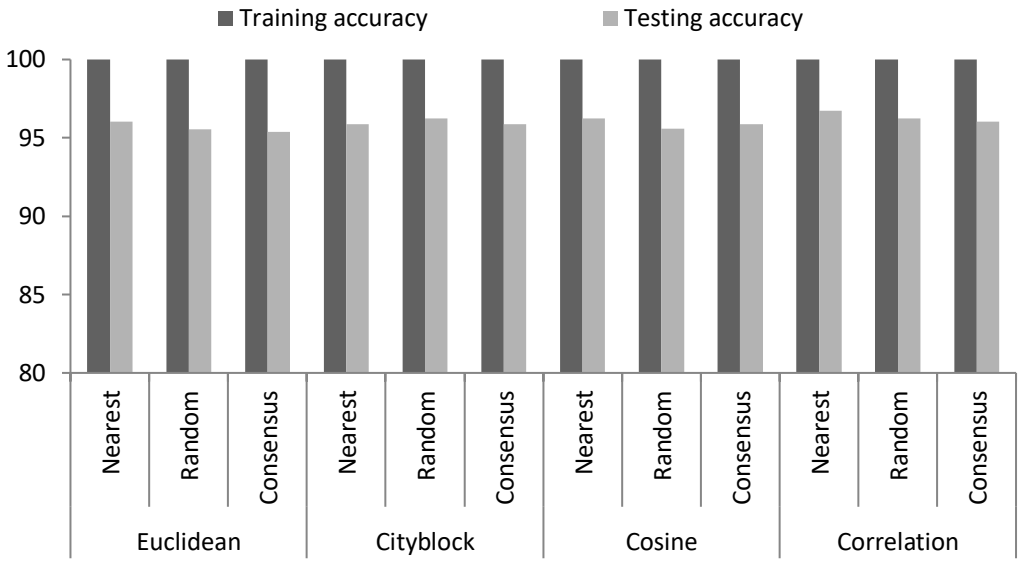

Figure 3.8: The comparative view of diagnostic accuracy rates of KNN based approaches

Figure 3.8 illustrates the comparative view of obtained diagnostic accuracy rates using KNN based approaches. KNN with euclidean distance metric and nearest rule had 100% (training) and 96.05% (testing) accuracy, KNN with euclidean distance metric and random rule had 100% (training) and 95.54% (testing) accuracy, KNN with euclidean distance metric and consensus rule had 100% (training) and 95.37% (testing) accuracy, KNN with cityblock distance metric and nearest rule had 100% (training) and 95.88% (testing) accuracy, KNN with cityblock distance metric and random rule had 100% (training) and 96.23% (testing) accuracy, KNN with cityblock distance metric and consensus rule had 100% (training) and 95.88% (testing) accuracy, KNN with cosine distance metric and nearest rule had 100% (training) and 96.23% (testing) accuracy, KNN with cosine distance metric and random rule had 100% (training) and 95.57% (testing) accuracy, KNN with cosine distance metric and consensus rule had 100% (training) and 95.88% (testing) accuracy, KNN with correlation distance metric and nearest rule had 100% (training) and 96.74% (testing) accuracy, KNN with correlation distance metric and random rule had 100% (training) and 96.23% (testing) accuracy and KNN with correlation distance metric and consensus rule had 100% (training) and 96.05% (testing) accuracy.

Table 3.2: The simulation results of integrated classification methods

Classification method		PCA-LDA	PCA-DLDA	PCA-QDA	PCA-DQDA	PCA-LSSVM	PCA-KNN
Accuracy	Training (%)	61.1	62.13	52.15	52.67	76.08	100
	Testing (%)	60.89	61.92	52.14	52.66	75.99	99.83
Sensitivity	Training (%)	78.92	77.11	94.58	96.39	26.51	100
	Testing (%)	78.44	76.65	94.61	96.41	26.35	100
Specificity	Training (%)	53.98	56.14	35.18	35.18	95.9	100
	Testing (%)	53.85	56.01	35.1	35.1	95.91	99.4
PPV	Training (%)	40.68	41.29	36.85	37.3	72.13	100
	Testing (%)	40.56	41.16	36.92	37.35	72.13	99.76
NPV	Training (%)	86.49	85.98	94.19	96.05	76.54	100
	Testing (%)	86.15	85.66	94.19	96.05	76.54	100

To select the most efficient liver diagnosis system, obtained results (accuracy, sensitivity, specificity, PPV and NPV) of all integrated methods were compared (Table 3.2). Prediction results of PCA-KNN are also compared to other liver classification methods mentioned in the literature. M. Abdar, et al. [117] obtained 93.75% accuracy using boosted C5.0; E. Elyan and M. M. Gaber [118] achieved 74.66% using RF-GA integration; F. Pourpanah, et al. [119] attained 80.82% using Q-learning multi-agent classifier system; S. Bashir, et al. [120] stated 72.7% using weighted multi-layer classifier ensemble framework; S. Bashir, et al. [121] mentioned 71.53% using a multi-layer classifier ensemble framework based on the optimal combinations; H. K. Sok, et al. [122] obtained 72.13% using multivariate alternating decision tree; P. C. Zou, et al. [123] et al. achieved 67.66% using NN with margin based metric learning algorithm; M. Abdar [124] attained 87.91% using C5.0 algorithm; R. M. O. Cruz, et al. [125] stated 69.40% using dynamic ensemble selection framework with meta-learning; H. K. Sok, et al. [126] mentioned 71.51% using sparse alternating decision tree; L. V. Utkin and A. I. Chekh [127] declared 72.2% using classification model with triangular kernel; Y. Xu, et al. [128] showed 72.74% using KNN-based weighted rough v-twin support vector machine; A. Yan, et al. [129] obtained 65.63% using enhanced case-based reasoning; H. Jin, et al. [130] achieved 65.3% using decision tree; and B. V. Ramana, et al. [131] attained 62.6% using support vector machines. It was found that PCA-KNN outperforms all other methods. This intelligent approach combines advantages of both PCA and KNN such as high classification rates, good generalization, plain structure and efficient problem solving ability through feature extraction. Generally, clinicians play the prime role in final judgment on patient's health condition but carrying out a resourceful diagnosis is an intricate job that requires enormous medical experience. Certainly, these computationally intelligent systems cannot replace physicians' role but may positively assist them in examining medical records by

acting as a second opinion. This study is also an effort in that direction which proposed a PCA-KNN based predictive method for the efficient and effective diagnosis of liver disease.

3.4 Conclusions

Development of ITs based computer-aided diagnostic systems is a productive work in clinical research. Implementation of these systems has contributed a major transformation in the field of information retrieval, and the medical domain has also been widely affected by this renovation. Number of authors have penned about the role of computational intelligence in medicine. Though disease diagnosis primarily relies on physician's clinical experience but computational intelligence does help in making precise judgments. Medicine field do have continuous advancements but diagnosing a disease is still challenging. Similarly, assessment of liver disease is also an intricate task. As a part of constant efforts for making liver diagnosis process well-organized and proficient, this chapter built a PCA-KNN based efficient diagnosis system with an inclusive analytic structure which boosts the prediction performance. The results showed an overall accuracy of 99.83%, sensitivity of 100%, specificity of 99.4%, positive predictive value of 99.76% and negative predictive value of 100%. An important observation from the experiment was that the key attributes extracted using principal component analysis were age, total bilirubin, alkaline phosphatase, and albumin and globulin ratio. These attributes were the most relevant indicators to identify presence of liver disease. K-nearest neighbor algorithm is a controlling learning technique that works well on basic recognition problems. It showed the capability of improving complex medical decisions based on training data by finding the k closest neighbors to a new instance. The prediction was carried out using a vast data of five hundred and eighty three samples from diverse patients. Experimental results confirmed the superiority of PCA-KNN to other diagnostic methods implemented in the study. False negative rates were reduced by dividing the data into training and testing. Thousands people lose their lives because of erroneous evaluation and inappropriate treatment as medical cases are still largely influenced by the subjectivity of clinicians. The proposed liver diagnostic system can be applied as a liver specialist assistant or as a model to train novice medical students. The system will also help physicians in evaluating complex cases that are otherwise hard to perceive.

The findings of the chapter have been published in the International Journal of Healthcare Information Systems and Informatics, Vol. 11, No. 4, pp. 56-69, 2016, IGI Global Publishing, DOI: 10.4018/IJHISI.2016100103, mentioned under list of publications at end of chapter 8.

Chapter 4

Implementation of Dimensionality Reduction Techniques and Classification Algorithms for the Prediction of Primary Biliary Cirrhosis Stages

This chapter is organized as follows: Section 4.1 introduces the chapter. Section 4.2 describes the methodologies deployed to classify primary biliary cirrhosis stages. Section 4.3 details the material used, discusses the experimental results and compare the performance of proposed intelligent integrated approach with other classification models. Finally, section 4.4 concludes the chapter.

4.1 Introduction

Primary biliary cirrhosis is a chronic, autoimmune and long-term liver disease in which the bile ducts in liver are destroyed due to inflammation. If the bile continues to remain in liver, it gradually converts into cirrhosis and then eventually to cancer. Bile, a liquid delivered in your liver, assumes a part in processing sustenance and frees your assortment of exhausted red platelets, cholesterol and poisons. At the point when bile pipes are harmed, as in PBC, destructive substances can develop in your liver and once in a while prompt to irreversible scarring of liver tissue [132]–[134]. PBC causes are not clear and it is viewed as an immune system illness, in which the body betrays its own particular cells. Medical experts assume it is activated by a mix of hereditary and ecological variables. Generally, half of the patients remain asymptotic but the side effects or signs of PBC can occur during any stage of the sickness and may incorporate weakness or reflect cholestasis, cirrhosis or hepatocellular brokenness. Weakness, pruritus, dry mouth and eyes are the underlying manifestations in greater than half of the patients. Other introductory indications incorporate an extended, firm, nontender liver; splenomegaly; right upper quadrant uneasiness; jaundice; xanthelasmas and hyperpigmentation with an approximate occurrence probability of twenty five, fifteen, ten, ten, ten and twenty five percent respectively. PBC for the most part grows gradually and prescription can moderate its movement, particularly if treatment starts early. Irritation of essential biliary cirrhosis starts when T lymphocytes begin amassing in the liver. Lymphocytes are white platelets that are a piece of insusceptible framework reaction. Irritation in the smallest conduits spreads, in time, and devastates close-by liver cells. As these cells are crushed, they're supplanted by scar tissue (fibrosis) that can add to cirrhosis.

Cirrhosis is scarring of liver tissue that makes it difficult for liver to do crucial capacities. PBC is a dynamic condition where the harm to liver can deteriorate with time. Its progression rate varies between individuals. If not treated, liver can get to be harmed to such a degree that it no longer works properly. This condition is represented as liver failure or fatal. There are prescriptions that can moderate the progress of damage and calm the irritation connected with it. In situations where there is broad liver harm, a liver transplant might be required. In this chapter, we have worked on a real-life biomedical dataset for the classification of PBC stages. In the given dataset [135], PBC is classified into four histologic stages include portal stage, periportal stage, septal stage and biliary cirrhosis stage which are represented as stage 1, stage 2, stage 3 and stage 4, respectively. Stage 4 represents the highest degree of damage, after which the condition can transform to more severe form of liver damage called as cirrhosis which is again a type of liver disease other than PBC. In stage 1, triads of ordinary size, portal inflammation, bile duct abnormalities, knobs loaded with an assortment of inflammatory cells are regularly recognized; in stage 2, periportal fibrosis is available with or without periportal irritation or prominent extension of the portal tracts; in stage 3, septal fibrosis with dynamic inflammatory, uninvolved paucicellular septa, or both are available; in stage 4, nodules with different degrees of inflammation are present. PBC disease is found in every five people out one lakh. It is widely present in women having a prevalence rate of ninety percent. Visibility of the disease is generally observed between the ages of forty and sixty years.

Medical interpretations from a collection of symptoms, risk factors, laboratory tests and other vital examination figures is a highly demanding task in liver disease diagnosis. The task becomes even more complex if the existing figures are fuzzy. It also stretches the decision time of physicians despite having a huge experience and if they are novice then it may take years for them to gain substantial expertise in analyzing uncertain health examination data of patients. Moreover, the accurate diagnosis is still not guaranteed as humans are prone to errors no matter whatever may the reason be like abundant clinical workload or a poor health. Medical data having variety of features increase the complexity of decision process. Prioritizing these features with respect to sickness is a prime task for effective assessment of patients' health. Hence, to interpret multifaceted data, to avoid clinical inexperience and to reduce the evaluation time, computer-aided prediction systems are built using a variety of intelligent techniques for diagnosis liver disease and its types. Applicability of these systems in liver disease evaluation has given a better facility to patients as well as doctors for correct and timely diagnosis [136]. Presently in medicine, there are number of computational models

available for assessing liver disease. However, a lot can be explored in building an intelligent computing model for classifying primary biliary cirrhosis as it has overlapping features within the stages. If a patient has an erroneous health evaluation due to ambiguity of medical attributes, then it would eventually lead to wrong treatment. Hence, it is a matter of prime concern to correctly classify PBC as this disease is also associated with other autoimmune disorders. Deployment of computational methods seems to be essential in order to interpret the complex structure of given dataset and to productively assist physicians in prediction process [4]. This chapter proposes two intelligent integrated models for the classification of primary biliary cirrhosis stages. These models are built using the integration of dimensionality reduction and classification algorithms. Dimensionality reduction constitutes of feature selection and feature extraction which are performed by using kullback-leibler divergence and principal component analysis techniques respectively. Classification algorithms deployed in this chapter are linear discriminant analysis, diagonal linear discriminant analysis, euclidean distance based k-nearest neighbor and least squares support vector machine. Performance of the proposed intelligent systems is evaluated on a real-life biomedical PBC dataset and is compared in terms of accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Experimental results showed that KLD-LSSVM and PCA-KNN based systems achieved better prediction outcomes than other computational methods in feature selection and feature extraction based methodologies respectively. The proposed models have shown robustness to the noisy data and also attained remarkable sensitivity and specificity.

4.2 Methodology

This section presents block diagrams of proposed intelligent integrated models implemented for classifying PBC stages. Figure 4.1 illustrates the overall structure design of feature selection based classification model where first step is loading the dataset, second is data preprocessing, third is data partitioning, fourth is performing feature selection, fifth is classifying the samples, and last step is performance evaluation and comparison which eventually decides the best classification model. In data preprocessing, each sample is symbolized as a vector of real numbers before giving input to a classifier. For instance, in selector field a sick class is represented by 1 and a healthy class is indicated with 0. The cross validation method divides the observations into training and testing set. Holdout cross validation method is used for data partitioning where seventy percent data is used for training and thirty percent data is used for testing. Feature selection is performed using kullback-

leibler divergence method which helps in finding the most discriminative features through ranking. Out of fifteen, first twelve features are selected, for giving input to the classifiers, based on their ranking. Intelligent integrated approaches implemented in feature selection based methodology are PCA-LDA, PCA-DLDA, PCA-KNN and PCA-LSSVM. Figure 4.2 depicts the overall structure design of feature extraction based classification model. It shows the classification of primary biliary cirrhosis stages using individual and integrated classification methods. K-fold cross validation method is used for data partitioning where k is having the value of five. Individual classifiers include linear discriminant analysis, diagonal linear discriminant analysis, and euclidean distance and nearest rule base k-nearest neighbor approach. Integrated models include combination of LDA, DLDA and KNN with principal component analysis. Prediction performances of the implemented models are compared for selecting the best classification models. The results of both the methodologies are computed using statistical parameters which include accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Description of kullback-leibler divergence method, principal component analysis technique and all the classification algorithms deployed to classify primary biliary cirrhosis stages are as follows.

Kullback-Leibler divergence distance method measures the difference between two probability distributions. It is also known as relative entropy and information divergence. The probability functions of two discrete distributions M and N are M_k and N_k respectively [137]. The relative entropy of M with respect to N is given in Eq. (4.1).

$$e = \sum_{k} M_k \log_2 \left(\frac{M_k}{N_k}\right) \tag{4.1}$$

Principal component analysis transforms large number of correlated variables into less numbers using various mathematical principles. These transformed variables are known as principal components which are here computed using vector space transformation method. This reduced dimension dataset with less variables helps to analyze patterns and outliers in the data which is very intricate without performing principal component analysis [138]–[140]. If there are p variables then p principal components are formed. The first principal component is formed by linear combination of variables that has largest variance and then other succeeding component are formed with the largest possible variance but the components are uncorrelated with previous components. Principal component is calculated by taking linear combination of an eigen vector of correlation matrix with variables. The eigen values in the

eigen vector represent the variance of each component. For instance, the given dataset have i-by-j data matrix D which consists of j number of i dimensional vectors $\overrightarrow{z_s} \in Q^i$.

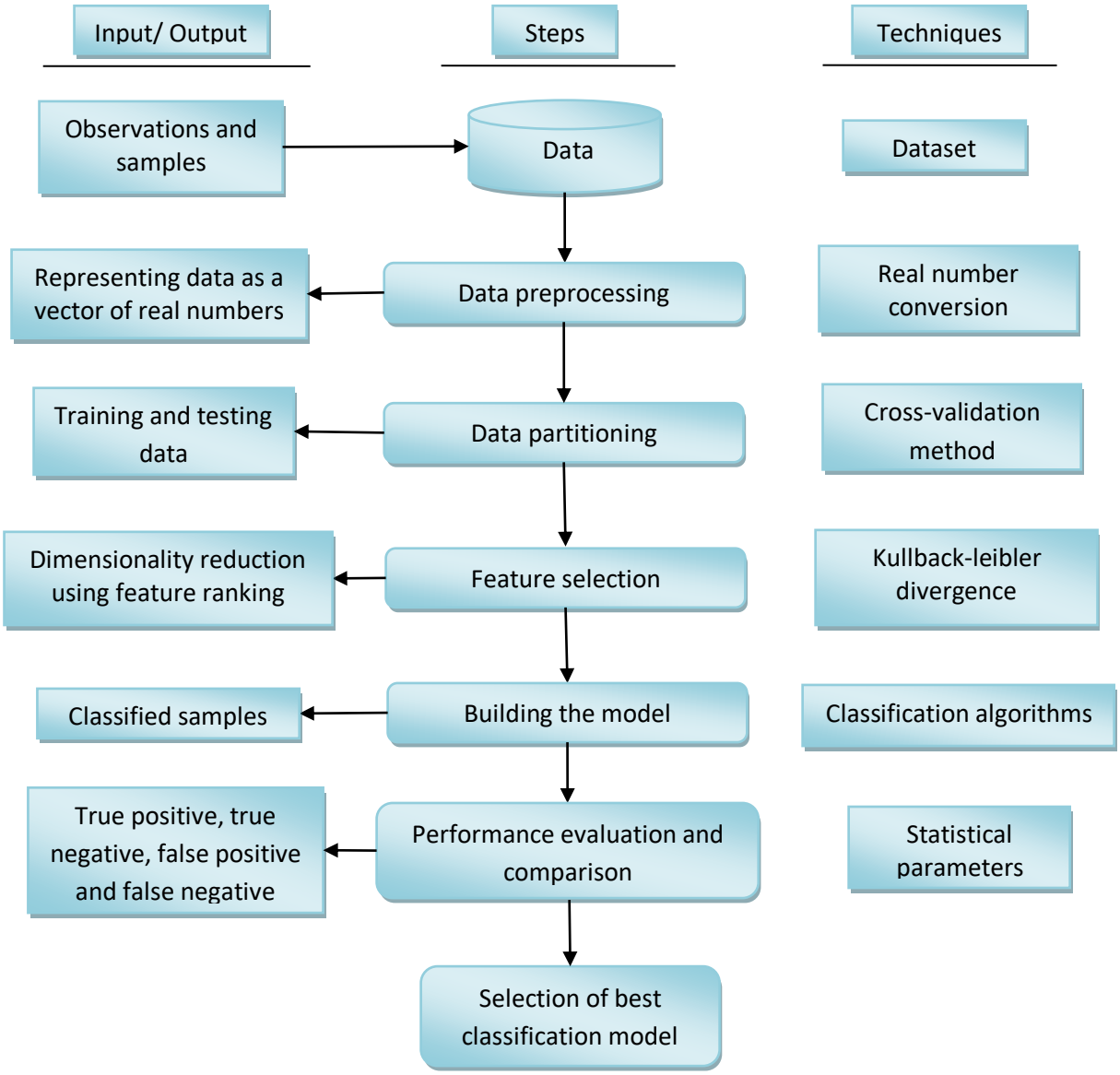

Figure 4.1: Block diagram of proposed feature selection based classification model for primary biliary cirrhosis stages

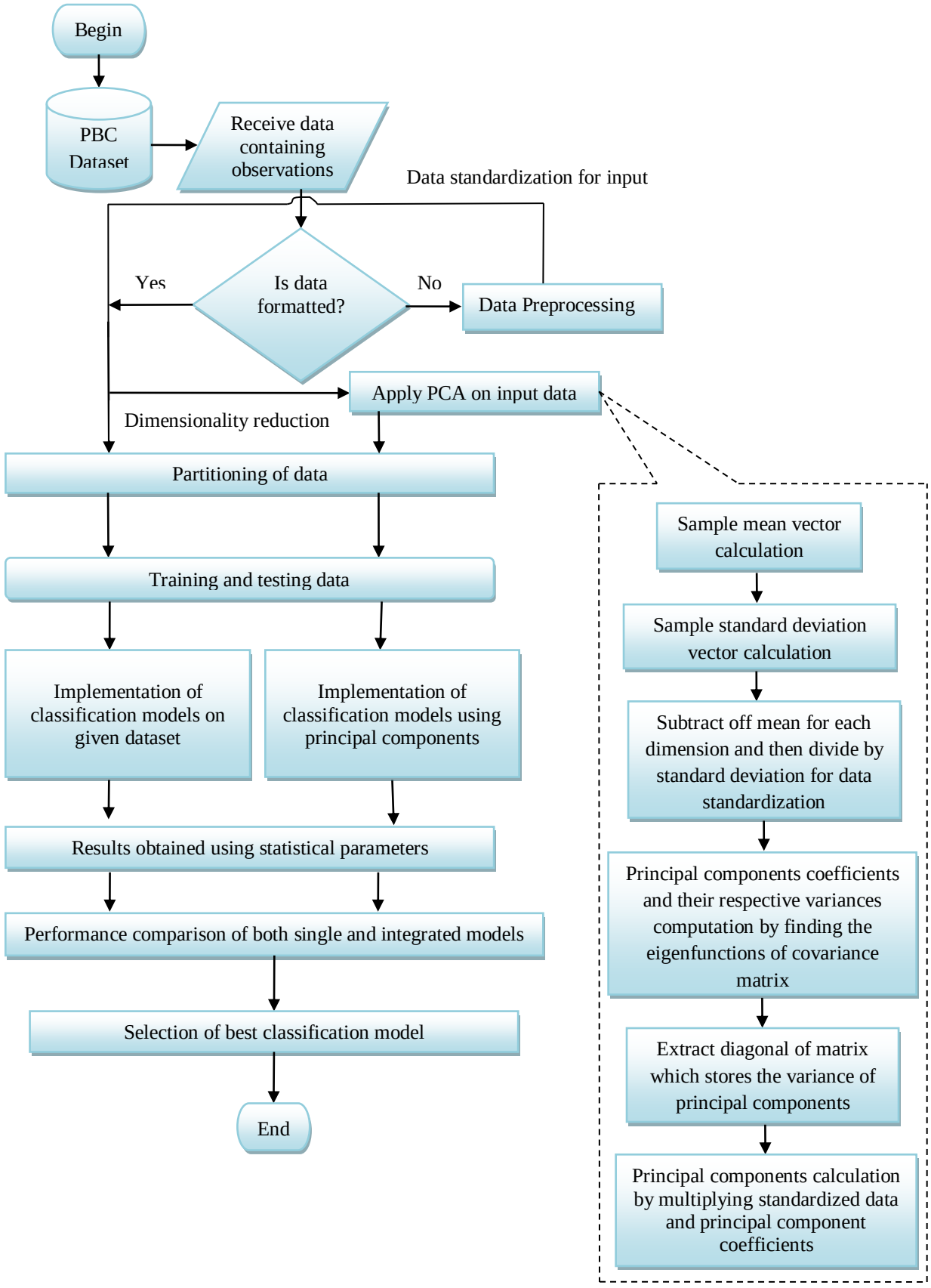

Figure 4.2: Block diagram of proposed feature extraction based classification model for primary biliary cirrhosis stages

Compute mean and covariance of data matrix

$T \in Q^{i \times i}$ is the covariance matrix of D which is defined as:

$$T = \frac{1}{j} \sum_{s=1}^{j} (\overrightarrow{z_s} - \bar{z}) \, (\overrightarrow{z_s} - \bar{z})^k \tag{4.2}$$

where $\overrightarrow{z_s} \in Q^i$ is the mean of each row of matrix D which is defined by:

$$\bar{z} = \frac{1}{j} \sum_{s=1}^{j} \overrightarrow{z_s} \tag{4.3}$$

Singular Vector Decomposition (SVD)

To extract principal components and directions, singular vector decomposition of T is:

$$T = W \, \Sigma \, Z^k \tag{4.4}$$

where $W \in Q^{j \times j}$, $\Sigma \in Q^{j \times i}$ and $Z \in Q^{i \times i}$. Here, we use the matrix $Z = [p_1 p_2 \dots \dots . . p_i]$ where vector $p_s \in Q^i$ represents principal component direction.

Projection

By multiplying a matrix B^k, the data matrix Y is projected into new matrix $X \in Q^{g \times i}$

$$X = B^k D \tag{4.5}$$

where $= [p_1 p_2 \dots \dots . . p_g]$, $g \leq i$. In order to perform projection of data matrix proper number of principal components g should be selected earlier.

Steps (V,G):

1) $\mathcal{E} \leftarrow mean\,(D)$ // mean computation
2) $V \leftarrow (D - \mathcal{E}1^k)^k \;\; (D - \mathcal{E}1^k)$ // covariance calculation
3) $\{\alpha_g, \mathcal{E}_g\} \leftarrow$ top G eigen values/eigen vectors of V
4) Return ($D - \mathcal{E}1$) W // projecting data using W.

Linear discriminant analysis works on the concept of searching for a linear combination of variables that best separates two classes. These variables are the predictors, and the classes are the actual targets in numerical form [105]–[107]. It works efficiently for disproportionate within-classes frequencies by maximizes the ratio of between-classes variance to within-classes variance for drawing decision region between the given classes. For example, let's assume that the dataset have *X* classes; class *j* mean vector is μ_j where *j=1, 2, . . X*; N_j indicates number of samples within class *j* where *j=1, 2, .. X*.

$$N = \sum_{j=0}^{X} N_j \tag{4.6}$$

$$M_a = \sum_{j=1}^{X} \sum_{i=1}^{N_i} (y_i - \mu_j)(y_i - \mu_j)^T \quad (4.7)$$

$$M_b = \sum_{j=1}^{X} (\mu_j - \mu)(\mu_j - \mu)^T \quad (4.8)$$

$$\mu = 1/X \sum_{j=1}^{X} \mu_j \quad (4.9)$$

where N is defined as the total number of samples, M_a is the within-class scatter matrix, M_b is the between-class scatter matrix and μ is the mean of entire dataset. On the other hand, DLDA is the extension of linear discriminant analysis where covariance matrices are assumed equal across groups.

K-nearest neighbor works on the concept of similarity measures. It preserves a range of cases and then a new case is classified based on various votes acquired from the neighbors. Its cost of the learning process is nearly zero. The simplest variation of KNN is default 1NN model having K=1, where the goal is to find a sample closest to the nearest neighbor y. The idea of single nearest neighbor approach is quite useful when data samples are large in number. The concept of 1NN can be extended to KNN in a way such that higher value of K reduces the effect of noise in given dataset. Instead of searching for a single sample, more samples are searched as nearest neighbor in order to remove the consequences of over-fitting [113], [141]. KNN imposes problems with the increase in number of dimensions. Its prediction accuracy is degraded with the increase in number of attributes. For maintaining the accuracy levels with the increase of attributes various dimensionality reduction techniques are applied in the pre-processing step of classification. Similarity measure in KNN is derived using variety of distance functions. Distance functions used in determining the case with maximum set of nearest neighbors are euclidean, correlation, cityblock, cosine, manhattan and minkowski out of which euclidean function is selected for the given dataset. Euclidean distance function is represented as mentioned in Eq (4.10).

$$Euclidean\ distance\ function = \sqrt{\sum_{j=1}^{k} (x_j - y_j)^2} \quad (4.10)$$

Support vector machine uses kernel functions to fit training data on a suitable hyperplane surface [110].. Setting the finest kernel parameters leads to constructive performance for the

problem. The hyperplane surface is established by support vector machine to separate a range of classes given in the problem. Maximum euclidean distance to the nearest point was used to attain the required hyperplane. Width of gaussian kernels and regularization factor were the two parameters included in LSSVM having a varied value between 1 and 100000 for the former and 0.1 and 10 for the later. Figure 4.3 presents the structure of a generic SVM that shows that the dataset contains two distinct classes like [+1, -1]. Assume that d data is contained in training data $(p_1, q_1), \ldots., (p_d, q_d),\ p \in R^d\ and\ q \in \{+1, -1\}$. Now, to find optimal hyperplane with maximum possible margin are formed with

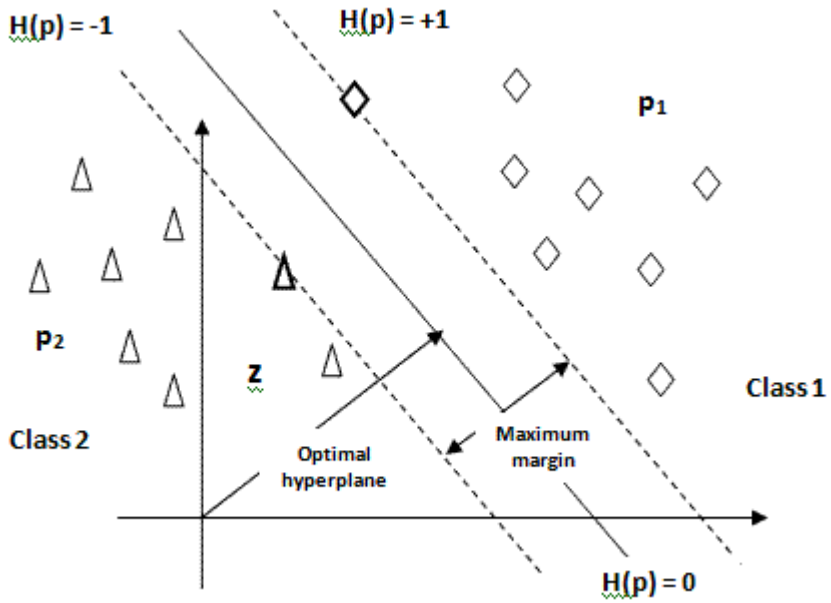

Figure 4.3: The structure of support vector machine

$$H(p) = (z.p) + z_0 \tag{4.11}$$

and the disparity for q = +1 and q = -1 is

$$q_i[(z.p_x) + z_0] \geq 1, x = 1, \ldots., d \tag{4.12}$$

This formula conferred by the data point p and q in equality condition are known as support vectors. The disparity abided by hyperplane margins are as follows.

$$\frac{ql \times H(pl)}{||z||} \geq \Gamma,\ \ l = 1, \ldots., d \tag{4.13}$$

Now, the equation (4.14) is resolute for dropping total number of solutions. The given formula is

$$\Gamma \times \left|\left|z\right|\right| = 1 \tag{4.14}$$

$$\frac{1}{2}||z||^2 \qquad (4.15),\ \textit{this is minimised subject to equation}\ (4.12)$$

In case of non-separable data slack variablesβ_i, are added in equation (4.12) and (4.15). The mentioned case then replaces the equation (4.12) and (4.15) with the new one as follows.

$$q_i[(z.p_x) + z_0] \geq 1 - \beta_x \qquad (4.16)$$

$$T\sum_{x=1}^{d}\beta_x + \frac{1}{2}\,||z||^2 \qquad (4.17)$$

This generic SVM functioning is productive in the linear classification of data but could not solve the nonlinear cases. To overcome this limitation kernel functions are being built which maps the given data into a kernel space. These kernel functions include linear kernels, quadratic kernels, radial basis function (RBF) kernels, polynomial kernels and multilayer perceptron (MLP) kernels. The study achieved best results with RBF kernels and least squares method in which the later was used to find the separating hyperplane [111]. RBF kernels is represented as $K(p,p') = \exp(-|\,|p-p'|\,|^2/\sigma^2)$. SVM uses a quadratic optimization problem for training which differs it from LSSVM in which a set of linear equations are used [112]. Like the concept defined in equation (4.16) and (4.17) for SVM, the same can be represented for LSSVM in the following form.

$$For\ x = 1, \ldots., d \qquad q_i[(z.p_x) + z_0] = 1 - \beta_x \qquad (4.18)$$

$$\frac{1}{2}\,||z||^2 + \frac{T}{2}\sum_{x=1}^{d}\beta_x^2 \qquad (4.19)$$

$$S(z,c,\alpha,\beta) = \frac{1}{2}\,||z||^2 + \frac{T}{2}\sum_{x=1}^{d}\beta_x^2 \; - \sum_{x=1}^{d}\alpha_x\,\{q_x[(z.p_x) + z_0] - 1 + \beta_x\} \qquad (4.20)$$

The α_x is knows as Lagrange multipliers which is an approach to find local maxima and minima for a function subject to a constraint. It should be positive in generic support vector machines structure but can be positive or negative in least squares support vector machine.

4.3 Results and Discussion

A mayo clinic dataset containing three hundred and fourteen medical records of primary biliary cirrhosis patients is taken for the study. In data preprocessing, records containing missing values are removed and a total of two hundred and seventy six samples are taken for experimentation. Features which seem to be non-informative for classification are also

removed. Finally sixteen features are incorporated which include age, serum albumin, alkaline phosphotase, presence of ascites, aspartate aminotransferase, serum bilirubin, serum cholesterol, urine copper, edema, hepatomegaly, platelet count, standardized blood clotting time, sex, spiders, triglycerides and histologic stage of disease. Each instance in the dataset represents information of a single male or female. For record, the data has thirty two male and two hundred and forty two female cases. Holdout and k-fold cross validation methods are used to validate the feature selection and feature extraction based proposed models respectively. In holdout, seventy percent data is used for training and thirty percent data is used for testing; and in k-fold, value of k is five. Description of features and their value ranges are given in Table 4.1. Obtained results are compared using accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) rates that are defined in Eq. (21), (22), (23), (24) and (25) respectively.

Table 4.1: Selected features of primary biliary cirrhosis dataset

Feature	Description	Values of features
Age	Age of the patient	in days
Gender	Gender of the patient	0 = female, 1 = male
Ascites	Presence of ascites	0 = no or 1 = yes
Hepato	Presence of hepatomegaly or enlarged liver	0 = no or 1 = yes
Spiders	Blood vessel malformations in the skin	0 = no or 1 = yes
Edema	Watery fluid collecting in the cavities or tissues	0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy
BIL	Serum bilirubin	in mg/dl
Chol	Serum cholesterol	in mg/dl
ALB	Serum albumin	in mg/dl
Copper	Urine copper	In mg/day
ALK	Alkaline phosphotase	in units/liter
AST	Aspartate aminotransferase	in units/ml
Trig	Triglycerides	in mg/dl
Platelet	Platelet count	per cubic ml/1000
Protime	Standardised blood clotting time	in seconds
Stages	Histologic stage of disease	1, 2, 3, 4

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.21}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.22}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.23}$$

$$PPV = \frac{TP}{TP + FP} \tag{4.24}$$

$$NPV = \frac{TN}{TN + FN} \tag{4.25}$$

where TN indicates true negative (normal people correctly recognized as normal), TP is true positive (diseased people correctly recognized as diseased), FN is false negative (diseased people incorrectly identified as normal), and FP expresses false positive (normal people incorrectly identified as diseased).

In feature selection based integrated models, firstly the features are ranked using KLD feature ranking technique for each PBC stage and then the first twelve features are given as input to the classifiers. Table 4.8 details the ranking of features with respect to the PBC stages. Figure 4.4, 4.5, 4.6 and 4.7 illustrate the comparative view of obtained accuracy rates for stage 1, 2, 3 and 4 respectively and Figure 4.8 presents the comparison of overall accuracy rates achieved by all integrated models. For stage 1, KLD-LDA had 78.87% (training) and 77.9% (testing) accuracy; KLD-DLDA had 72.16% (training) and 72.1% (testing) accuracy; KLD-KNN had 100% (training) and 98.19% (testing) accuracy; and KLD-LSSVM had 97.42% (training) and 97.1% (testing) accuracy. For stage 2, KLD-LDA had 59.79% (training) and 61.23% (testing) accuracy; KLD-DLDA had 60.31% (training) and 63.41% (testing) accuracy; KLD-KNN had 100% (training) and 91.3% (testing) accuracy; and KLD-LSSVM had 96.39% (training) and 91.3% (testing) accuracy. For stage 3, KLD-LDA had 67.53% (training) and 64.86% (testing) accuracy; KLD-DLDA had 61.86% (training) and 59.42% (testing) accuracy; KLD-KNN had 100% (training) and 86.59% (testing) accuracy; and KLD-LSSVM had 97.42% (training) and 86.59% (testing) accuracy. For stage 4, KLD-LDA had 77.84% (training) and 78.62% (testing) accuracy; KLD-DLDA had 76.8% (training) and 78.62% (testing) accuracy; KLD-KNN had 100% (training) and 86.59% (testing) accuracy; and KLD-LSSVM had 97.42% (training) and 88.77% (testing) accuracy. The overall accuracy rates of KLD-LDA were 71.01% (training) and 70.65% (testing); KLD-DLDA were 67.78% (training) and 68.39% (testing); KLD-KNN were 100% (training) and 90.67% (testing); and KLD-LSSVM were 97.16% (training) and 90.94% (testing).

Figure 4.9, 4.10, 4.11 and 4.12 illustrate the comparative view of obtained sensitivity rates for stage 1, 2, 3 and 4 respectively and Figure 4.13 presents the comparison of overall

sensitivity rates by all integrated models. For stage 1, KLD-LDA had 78.92% (training) and 78.03% (testing) sensitivity; KLD-DLDA had 71.35% (training) and 71.21% (testing) sensitivity; KLD-KNN had 100% (training) and 99.24% (testing) sensitivity; and KLD-LSSVM had 100% (training) and 100% (testing) sensitivity. For stage 2, KLD-LDA had 56.58% (training) and 58.06% (testing) sensitivity; KLD-DLDA had 55.26% (training) and 58.53% (testing) sensitivity; KLD-KNN had 100% (training) and 94.47% (testing) sensitivity; and KLD-LSSVM had 100% (training) and 100% (testing) sensitivity. For stage 3, KLD-LDA had 67.24% (training) and 62.42% (testing) sensitivity; KLD-DLDA had 53.45% (training) and 49.09% (testing) sensitivity; KLD-KNN had 100% (training) and 87.88% (testing) sensitivity; and KLD-LSSVM had 97.41% (training) and 91.52% (testing) sensitivity. For stage 4, KLD-LDA had 79.69% (training) and 80.22% (testing) sensitivity; KLD-DLDA had 85.16% (training) and 85.71% (testing) sensitivity; KLD-KNN had 100% (training) and 90.66% (testing) sensitivity; and KLD-LSSVM had 100% (training) and 97.8% (testing) sensitivity. The overall sensitivity rates of KLD-LDA were 70.61% (training) and 69.68% (testing); KLD-DLDA were 66.31% (training) and 66.14% (testing); KLD-KNN were 100% (training) and 93.06% (testing); and KLD-LSSVM were 99.35% (training) and 97.33% (testing).

Figure 4.14, 4.15, 4.16 and 4.17 illustrate the comparative view of obtained specificity rates for stage 1, 2, 3 and 4 respectively and Figure 4.18 presents the comparison of overall specificity rates by all integrated models. For stage 1, KLD-LDA had 77.78% (training) and 75% (testing) specificity; KLD-DLDA had 88.89% (training) and 91.67% (testing) specificity; KLD-KNN had 100% (training) and 75% (testing) specificity; and KLD-LSSVM had 44.44% (training) and 33.33% (testing) specificity. For stage 2, KLD-LDA had 71.43% (training) and 72.88% (testing) specificity; KLD-DLDA had 78.57% (training) and 81.36% (testing) specificity; KLD-KNN had 100% (training) and 79.66% (testing) specificity; and KLD-LSSVM had 83.33% (training) and 59.32% (testing) specificity. For stage 3, KLD-LDA had 67.95% (training) and 68.47% (testing) specificity; KLD-DLDA had 74.36% (training) and 74.77% (testing) specificity; KLD-KNN had 100% (training) and 84.68% (testing) specificity; and KLD-LSSVM had 97.44% (training) and 79.28% (testing) specificity. For stage 4, KLD-LDA had 74.24% (training) and 75.53% (testing) specificity; KLD-DLDA had 60.61% (training) and 64.89% (testing) specificity; KLD-KNN had 100% (training) and 78.72% (testing) specificity; and KLD-LSSVM had 92.42% (training) and 71.28% (testing) specificity. The overall specificity rates of KLD-LDA were 74.35% (training) and 72.97%

(testing); KLD-DLDA were 75.61% (training) and 78.17% (testing); KLD-KNN were 100% (training) and 79.52% (testing); and KLD-LSSVM were 79.41% (training) and 60.8% (testing).

Figure 4.19, 4.20, 4.21 and 4.22 illustrate the comparative view of obtained PPV rates for stage 1, 2, 3 and 4 respectively and Figure 4.23 presents the comparison of overall PPV rates by all integrated models. For stage 1, KLD-LDA had 98.65% (training) and 98.56% (testing) PPV; KLD-DLDA had 99.25% (training) and 99.47% (testing) PPV; KLD-KNN had 100% (training) and 98.87% (testing) PPV; and KLD-LSSVM had 97.37% (training) and 97.06% (testing) PPV. For stage 2, KLD-LDA had 87.76% (training) and 88.73% (testing) PPV; KLD-DLDA had 90.32% (training) and 92.03% (testing) PPV; KLD-KNN had 100% (training) and 94.47% (testing) PPV; and KLD-LSSVM had 95.6% (training) and 90.04% (testing) PPV. For stage 3, KLD-LDA had 75.73% (training) and 74.64% (testing) PPV; KLD-DLDA had 75.61% (training) and 74.31% (testing) PPV; KLD-KNN had 100% (training) and 89.51% (testing) PPV; and KLD-LSSVM had 98.26% (training) and 86.78% (testing) PPV. For stage 4, KLD-LDA had 85.71% (training) and 86.39% (testing) PPV; KLD-DLDA had 80.74% (training) and 82.54% (testing) PPV; KLD-KNN had 100% (training) and 89.19% (testing) PPV; and KLD-LSSVM had 96.24% (training) and 86.83% (testing) PPV. The overall PPV rates of KLD-LDA were 86.96% (training) and 87.08% (testing); KLD-DLDA were 86.48% (training) and 87.09% (testing); KLD-KNN were 100% (training) and 93.01% (testing); and KLD-LSSVM were 96.87% (training) and 90.18% (testing).

Figure 4.24, 4.25, 4.26 and 4.27 illustrate the comparative view of obtained NPV rates for stage 1, 2, 3 and 4 respectively and Figure 4.28 presents the comparison of overall NPV rates by all integrated models. For stage 1, KLD-LDA had 15.22% (training) and 13.43% (testing) NPV; KLD-DLDA had 13.11% (training) and 12.64% (testing) NPV; KLD-KNN had 100% (training) and 81.82% (testing) NPV; and KLD-LSSVM had 100% (training) and 100% (testing) NPV. For stage 2, KLD-LDA had 31.25% (training) and 32.09% (testing) NPV; KLD-DLDA had 32.67% (training) and 34.78% (testing) NPV; KLD-KNN had 100% (training) and 79.66% (testing) NPV; and KLD-LSSVM had 100% (training) and 100% (testing) NPV. For stage 3, KLD-LDA had 58.24% (training) and 55.07% (testing) NPV; KLD-DLDA had 51.79% (training) and 49.7% (testing) NPV; KLD-KNN had 100% (training) and 82.46% (testing) NPV; and KLD-LSSVM had 96.2% (training) and 86.27%

(testing) NPV. For stage 4, KLD-LDA had 65.33% (training) and 66.36% (testing) NPV; KLD-DLDA had 67.8% (training) and 70.11% (testing) NPV; KLD-KNN had 100% (training) and 81.32% (testing) NPV; and KLD-LSSVM had 100% (training) and 94.37% (testing) NPV. The overall NPV rates of KLD-LDA were 42.51% (training) and 41.74% (testing); KLD-DLDA were 41.34% (training) and 41.81% (testing); KLD-KNN were 100% (training) and 81.32% (testing); and KLD-LSSVM were 99.05% (training) and 95.16% (testing).

In feature extraction based integrated models, firstly the features are extracted using principal component analysis and then the first twelve features are taken based on their score values and were given as input to classifiers. The objective of this methodology is to compare the performance of individual and integrated models; and to prove the credibility of dimensionality reduction technique in enhancing the prediction results. Figure 4.29, 4.30, 4.31, 4.32 and 4.33 illustrate the comparative view of obtained accuracy, sensitivity, specificity, PPV and NPV rates of individual and integrated models respectively. In individual classifiers, LDA had 52.94% (training) and 51.45% (testing) accuracy, 77.78% (training) and 75% (testing) sensitivity, 83.96% (training) and 82.95% (testing) specificity, 17.07% (training) and 16.67% (testing) PPV; and 98.89% (training) and 98.65% (testing) NPV. DLDA had 51.11% (training) and 50.36% (testing) accuracy, 76.19% (training) and 75% (testing) sensitivity, 83.82% (training) and 83.71% (testing) specificity, 17.2% (training) and 17.31% (testing) PPV; and 98.76% (training) and 98.66% (testing) NPV. KNN had 100% (training) and 88.41% (testing) accuracy, 100% (training) and 83.33% (testing) sensitivity, 100% (training) and 98.48% (testing) specificity, 100% (training) and 71.43% (testing) PPV; and 100% (training) and 99.24% (testing) NPV. In integrated systems, PCA-LDA had 54.75% (training) and 53.62% (testing) accuracy, 80% (training) and 66.67% (testing) sensitivity, 89.1% (training) and 88.64% (testing) specificity, 25.81% (training) and 21.05% (testing) PPV; and 98.95% (training) and 98.32% (testing) NPV. PCA-DLDA had 55.2% (training) and 55.07% (testing) accuracy, 50% (training) and 41.67% (testing) sensitivity, 89.57% (training) and 89.39% (testing) specificity, 18.52% (training) and 15.15% (testing) PPV; and 97.42% (training) and 97.12% (testing) NPV. PCA-KNN had 100% (training) and 91.3% (testing) accuracy, 100% (training) and 83.33% (testing) sensitivity, 100% (training) and 99.62% (testing) specificity, 100% (training) and 90.91% (testing) PPV; and 100% (training) and 99.25% (testing) NPV.

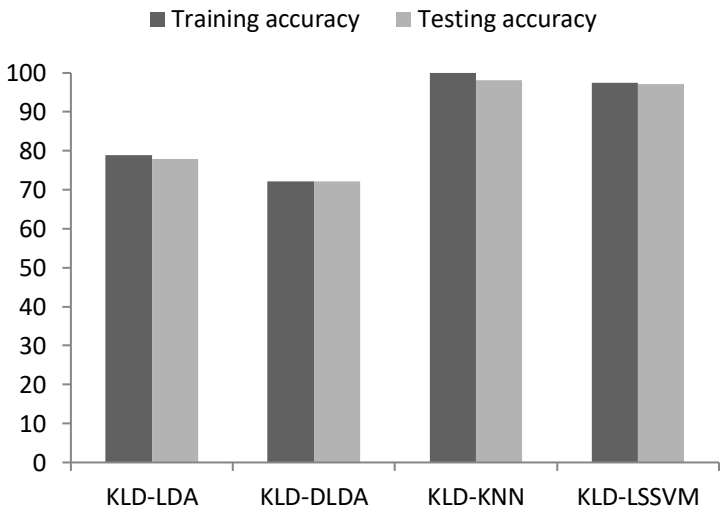

Figure 4.4: Comparative view of obtained accuracy rates of integrated models for stage 1

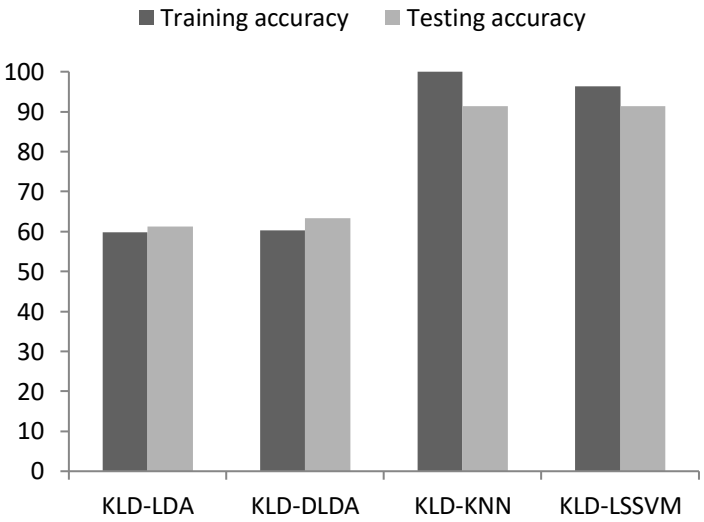

Figure 4.5: Comparative view of obtained accuracy rates of integrated models for stage 2

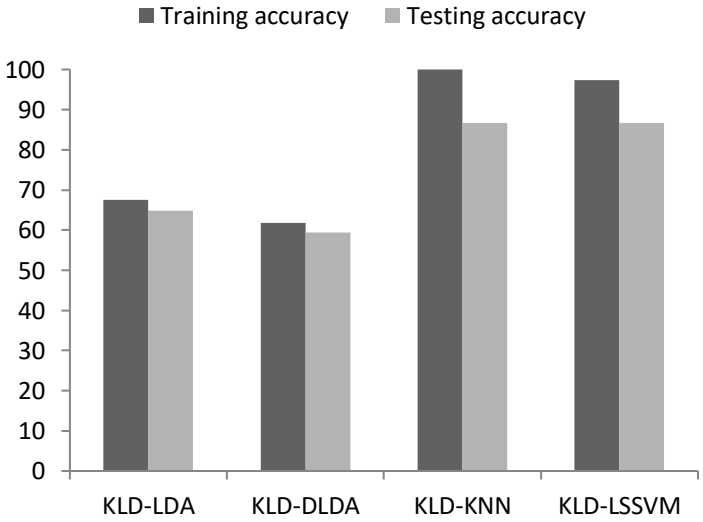

Figure 4.6: Comparative view of obtained accuracy rates of integrated models for stage 3

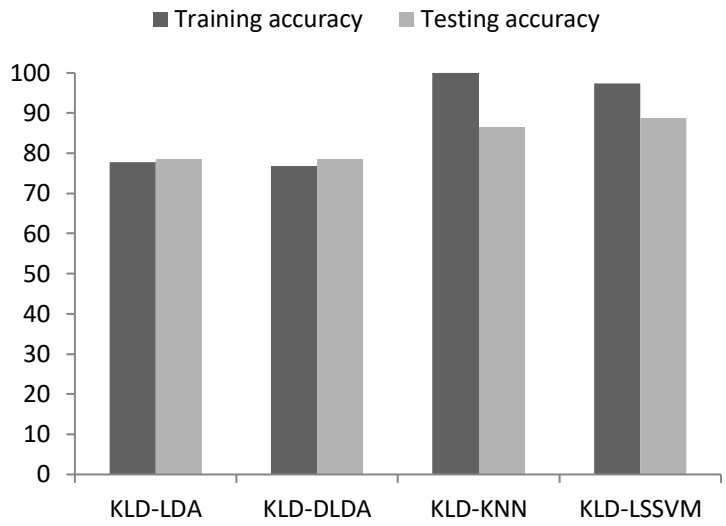

Figure 4.7: Comparative view of obtained accuracy rates of integrated models for stage 4

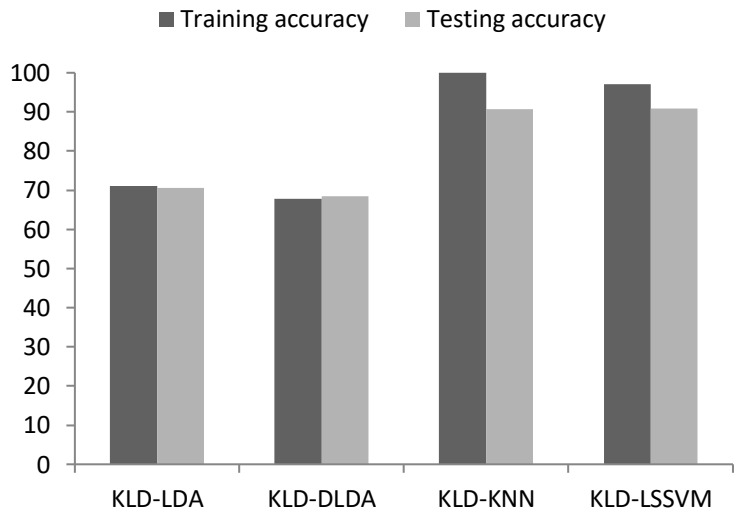

Figure 4.8: Comparative view of obtained overall accuracy rates of integrated models

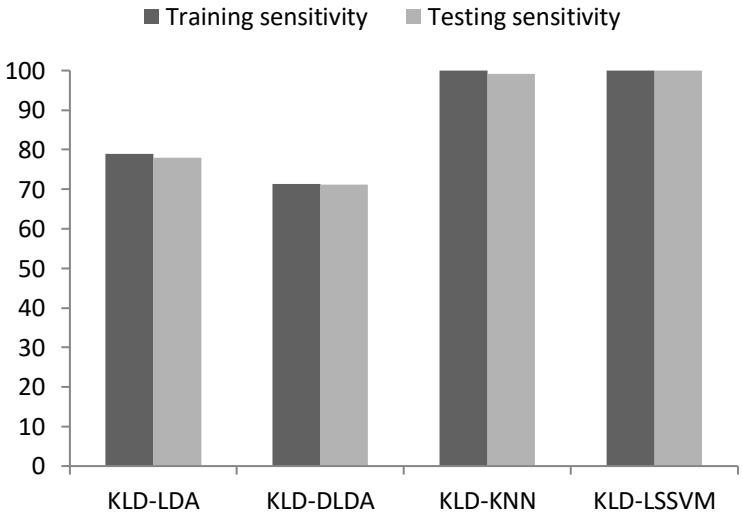

Figure 4.9: Comparative view of obtained sensitivity rates of integrated models for stage 1

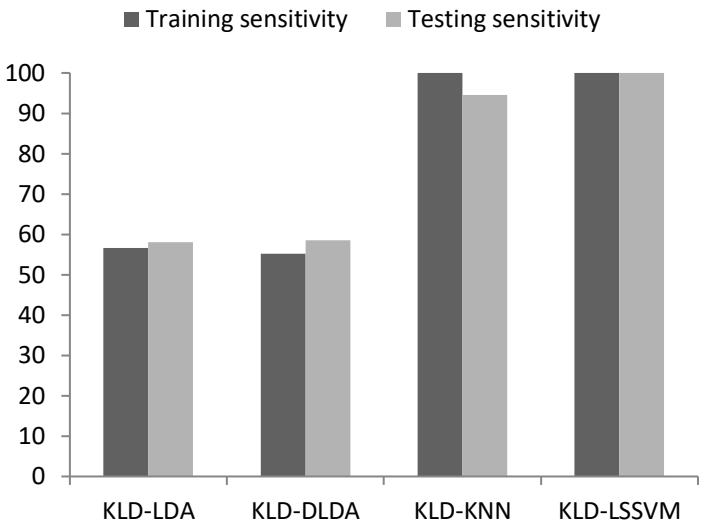

Figure 4.10: Comparative view of obtained sensitivity rates of integrated models for stage 2

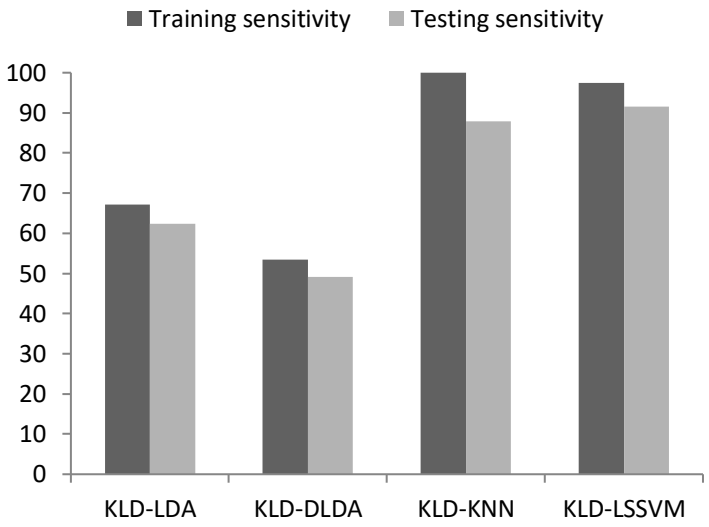

Figure 4.11: Comparative view of obtained sensitivity rates of integrated models for stage 3

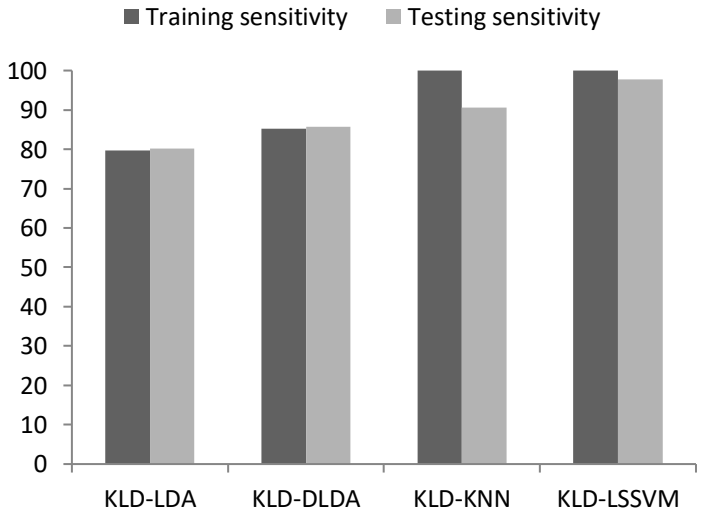

Figure 4.12: Comparative view of obtained sensitivity rates of integrated models for stage 4

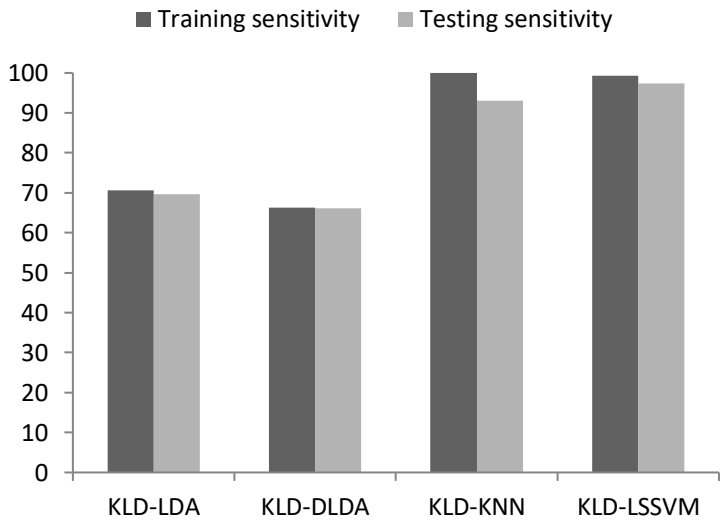

Figure 4.13: Comparative view of obtained overall sensitivity rates of integrated models

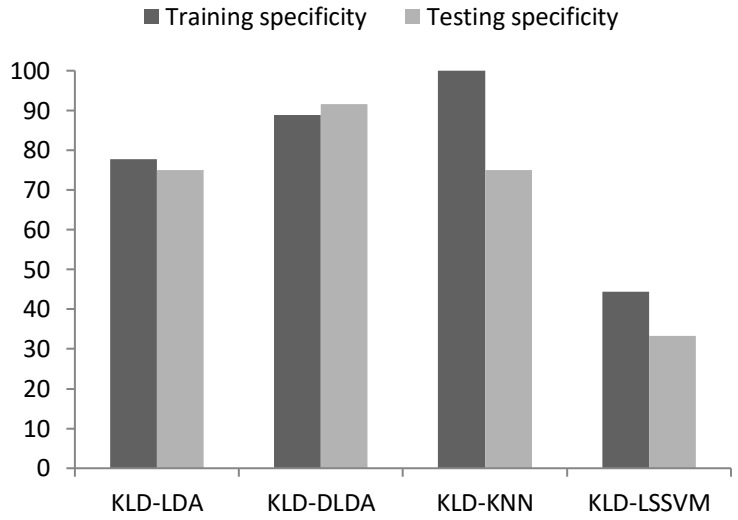

Figure 4.14: Comparative view of obtained specificity rates of integrated models for stage 1

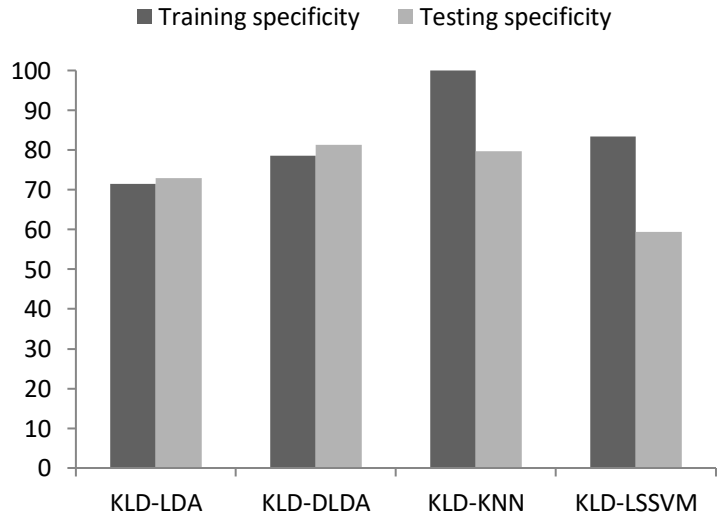

Figure 4.15: Comparative view of obtained specificity rates of integrated models for stage 2

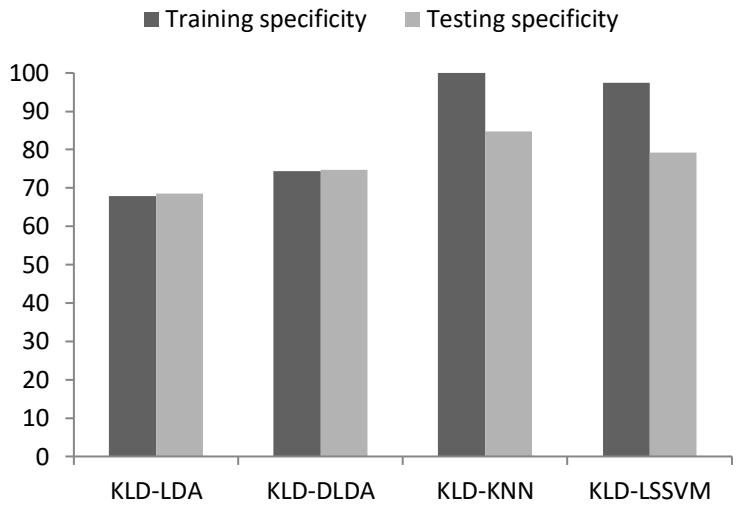

Figure 4.16: Comparative view of obtained specificity rates of integrated models for stage 3

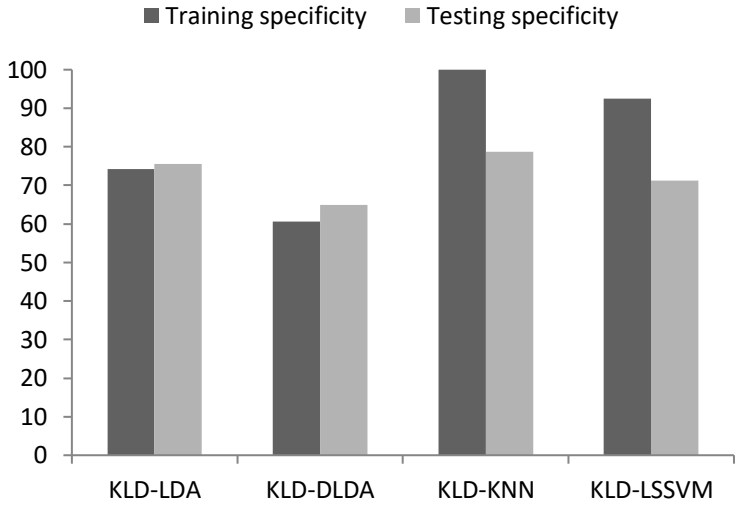

Figure 4.17: Comparative view of obtained specificity rates of integrated models for stage 4

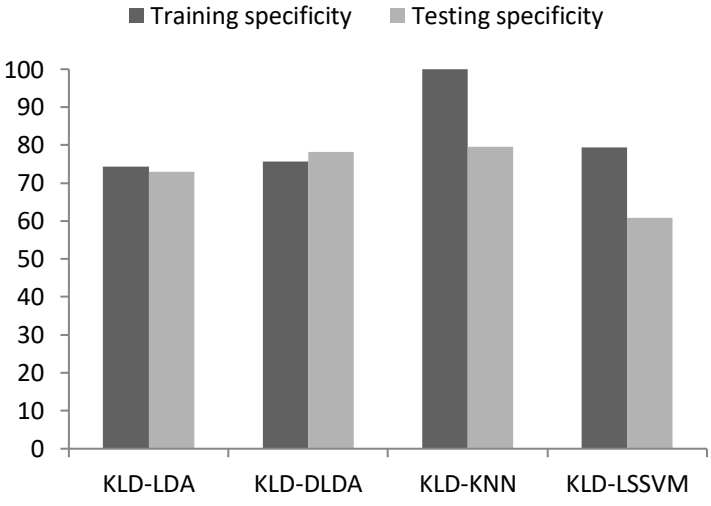

Figure 4.18: Comparative view of obtained overall specificity rates of integrated models

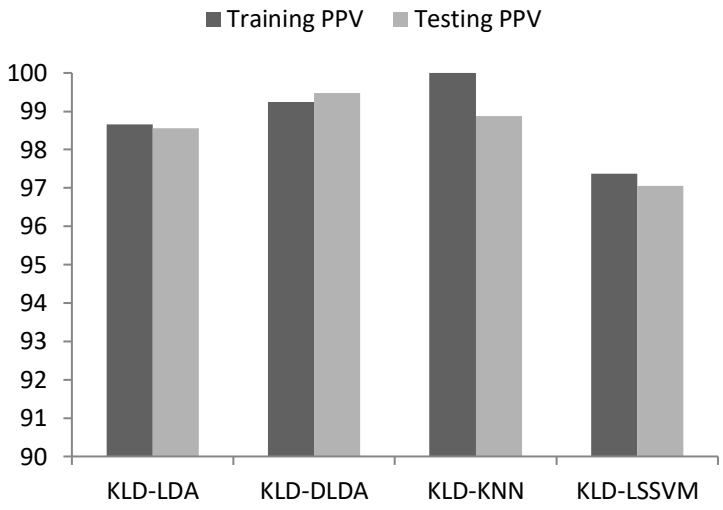

Figure 4.19: Comparative view of obtained PPV rates of integrated models for stage 1

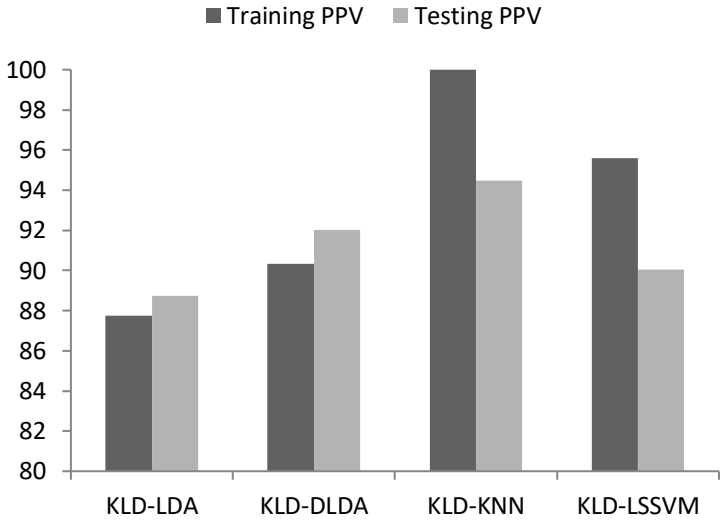

Figure 4.20: Comparative view of obtained PPV rates of integrated models for stage 2

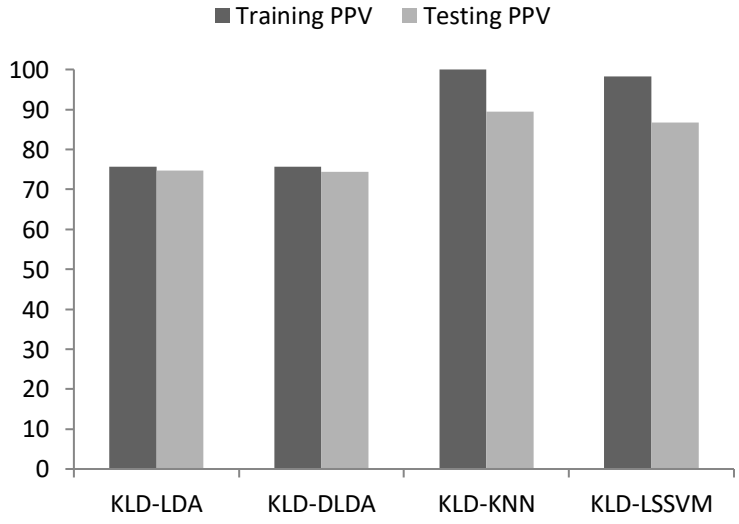

Figure 4.21: Comparative view of obtained PPV rates of integrated models for stage 3

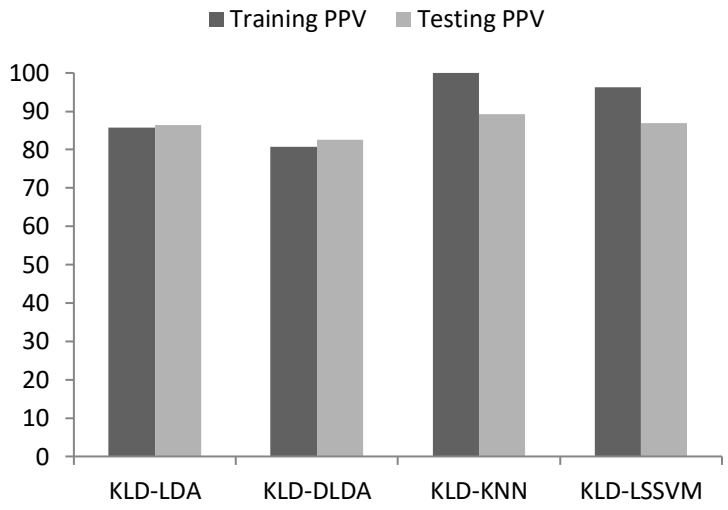

Figure 4.22: Comparative view of obtained PPV rates of integrated models for stage 4

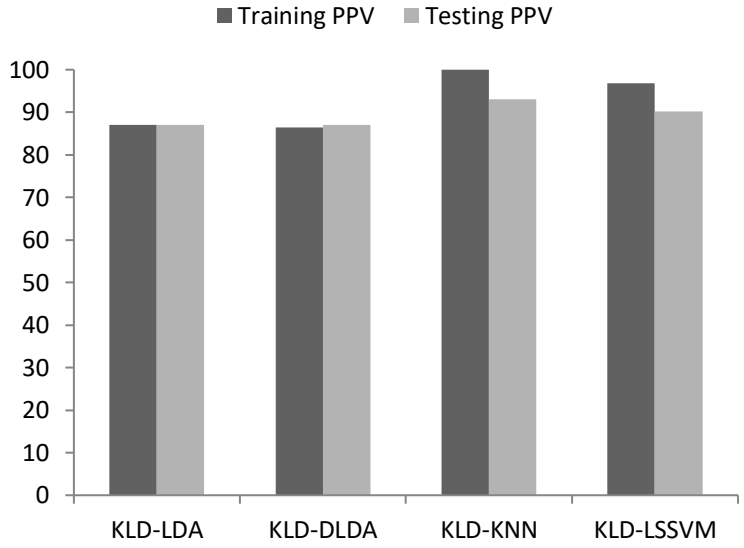

Figure 4.23: Comparative view of obtained overall PPV rates of integrated models

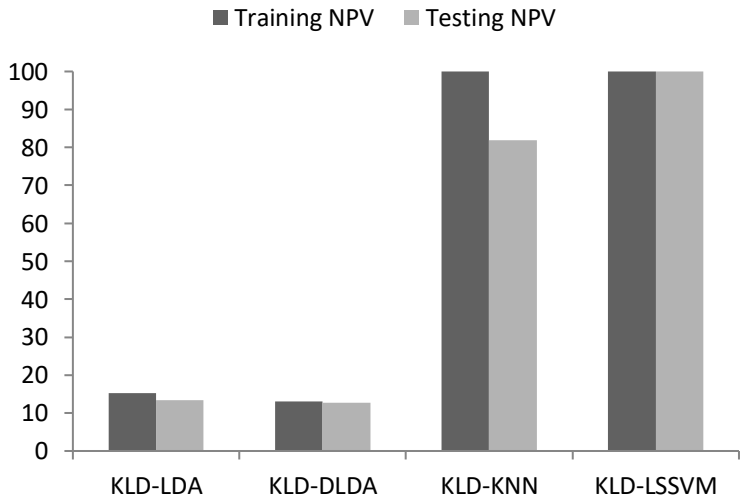

Figure 4.24: Comparative view of obtained NPV rates of integrated models for stage 1

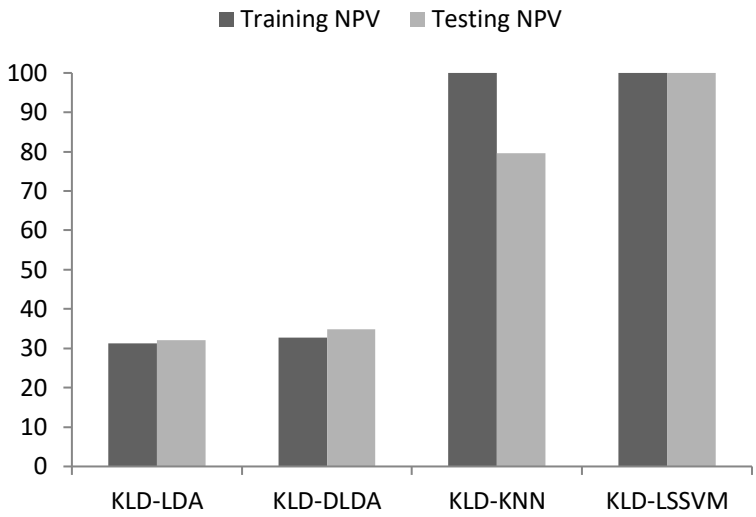

Figure 4.25: Comparative view of obtained NPV rates of integrated models for stage 2

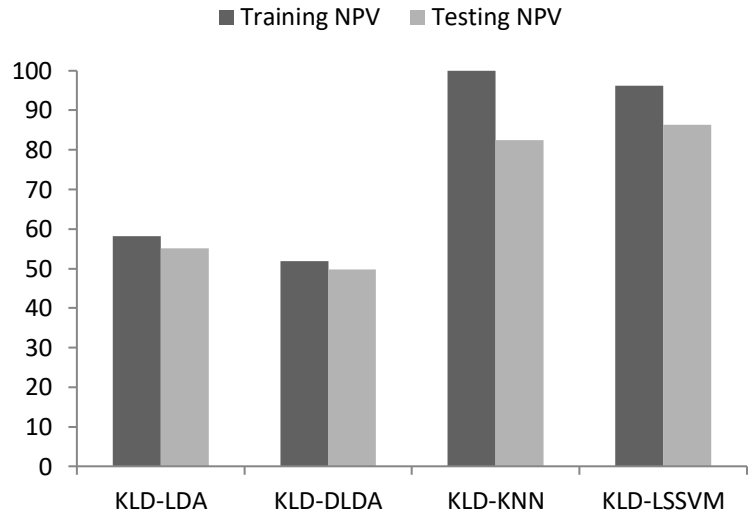

Figure 4.26: Comparative view of obtained NPV rates of integrated models for stage 3

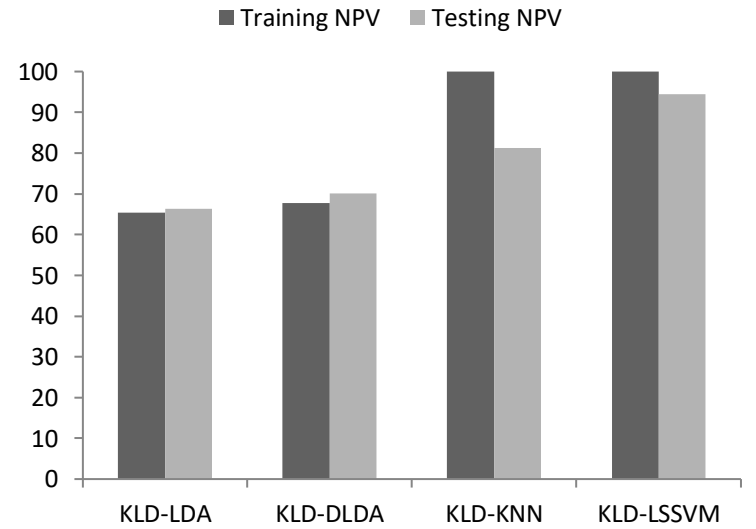

Figure 4.27: Comparative view of obtained NPV rates of integrated models for stage 4

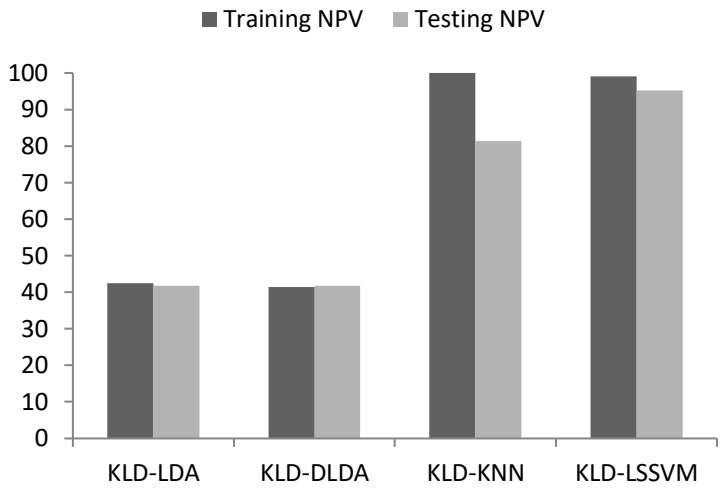

Figure 4.28: Comparative view of obtained overall NPV rates of integrated models

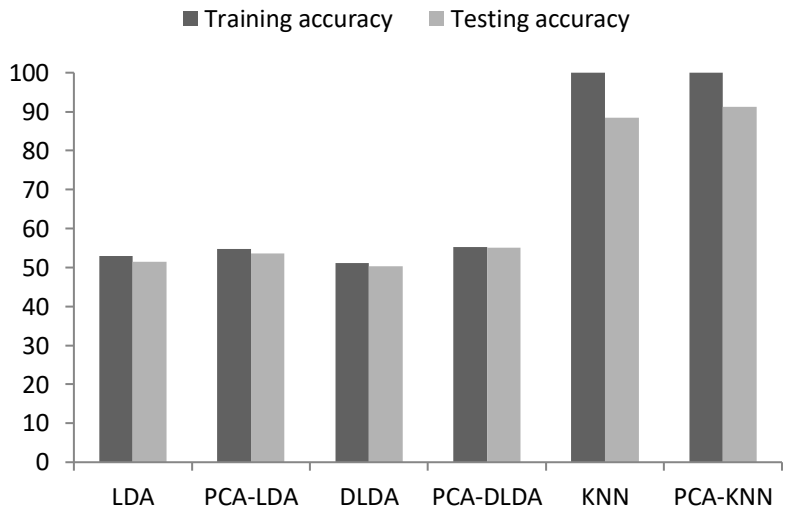

Figure 4.29: Comparative view of obtained accuracy rates of individual and integrated models

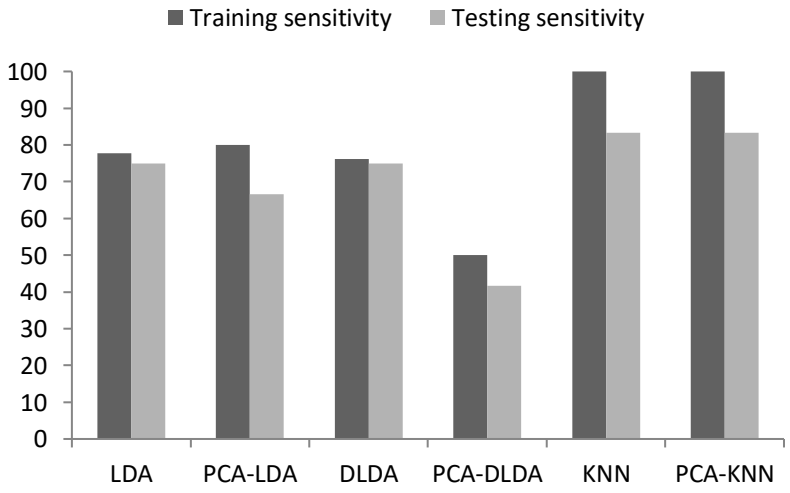

Figure 4.30: Comparative view of obtained sensitivity rates of individual and integrated models

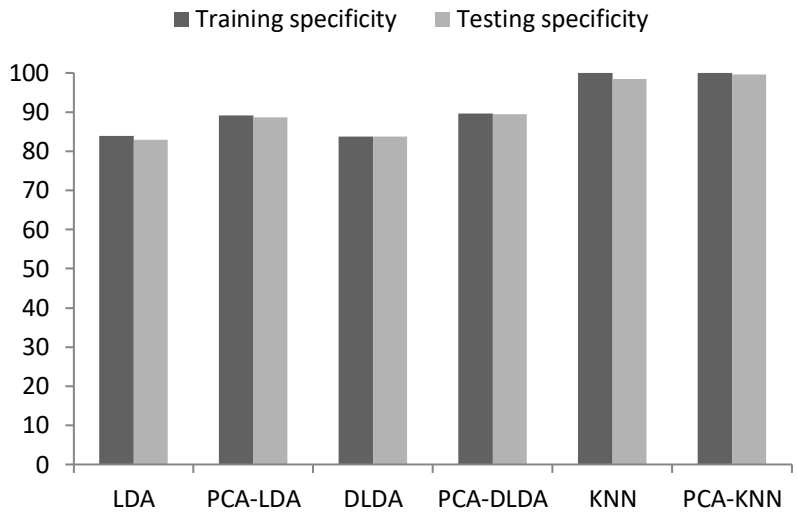

Figure 4.31: Comparative view of obtained specificity rates of individual and integrated models

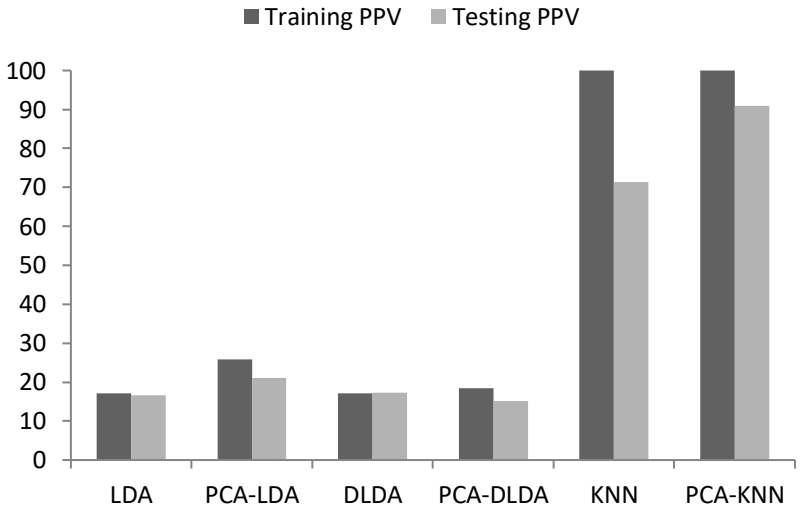

Figure 4.32: Comparative view of obtained PPV rates of individual and integrated models

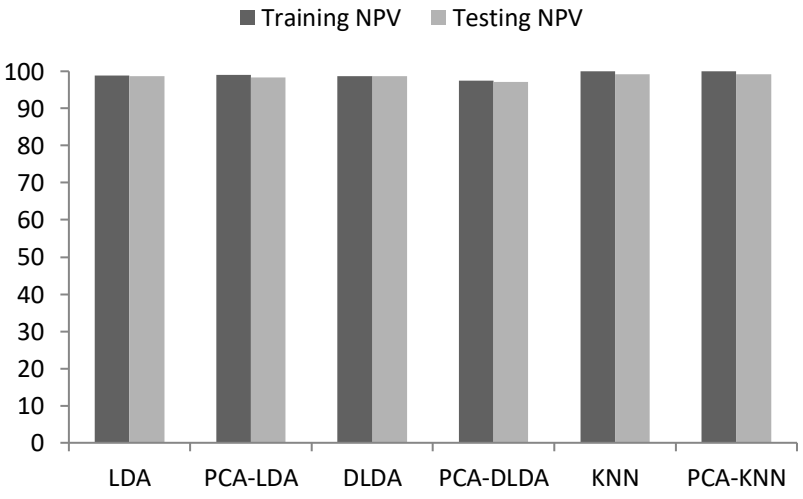

Figure 4.33: Comparative view of obtained NPV rates of individual and integrated models

Table 4.2: The achieved accuracy results of integrated computational methods

Integrated model		KLD-LDA	KLD-DLDA	KLD-KNN	KLD-LSSVM
Stage 1 Accuracy	Training (%)	78.87	72.16	100	97.42
	Testing (%)	77.90	72.10	98.19	97.10
Stage 2 Accuracy	Training (%)	59.79	60.31	100	96.39
	Testing (%)	61.23	63.41	91.30	91.30
Stage 3 Accuracy	Training (%)	67.53	61.86	100	97.42
	Testing (%)	64.86	59.42	86.59	86.59
Stage 4 Accuracy	Training (%)	77.84	76.80	100	97.42
	Testing (%)	78.62	78.62	86.59	88.77
Overall accuracy	Training (%)	71.01	67.78	100	97.16
	Testing (%)	70.65	68.39	90.67	90.94

Table 4.3: The achieved sensitivity results of integrated computational methods

Integrated model		KLD-LDA	KLD-DLDA	KLD-KNN	KLD-LSSVM
Stage 1 sensitivity	Training (%)	78.92	71.35	100	100
	Testing (%)	78.03	71.21	99.24	100
Stage 2 sensitivity	Training (%)	56.58	55.26	100	100
	Testing (%)	58.06	58.53	94.47	100
Stage 3 sensitivity	Training (%)	67.24	53.45	100	97.41
	Testing (%)	62.42	49.09	87.88	91.52
Stage 4 sensitivity	Training (%)	79.69	85.16	100	100
	Testing (%)	80.22	85.71	90.66	97.8
Overall sensitivity	Training (%)	70.61	66.31	100	99.35
	Testing (%)	69.68	66.14	93.06	97.33

Table 4.4: The achieved specificity results of integrated computational methods

Integrated model		KLD-LDA	KLD-DLDA	KLD-KNN	KLD-LSSVM
Stage 1 specificity	Training (%)	77.78	88.89	100	44.44
	Testing (%)	75	91.67	75	33.33
Stage 2 specificity	Training (%)	71.43	78.57	100	83.33
	Testing (%)	72.88	81.36	79.66	59.32
Stage 3 specificity	Training (%)	67.95	74.36	100	97.44
	Testing (%)	68.47	74.77	84.68	79.28
Stage 4 specificity	Training (%)	74.24	60.61	100	92.42
	Testing (%)	75.53	64.89	78.72	71.28
Overall specificity	Training (%)	74.35	75.61	100	79.41
	Testing (%)	72.97	78.17	79.52	60.8

Table 4.5: The achieved PPV results of integrated computational methods

Integrated model		KLD-LDA	KLD-DLDA	KLD-KNN	KLD-LSSVM
Stage 1 PPV	Training (%)	98.65	99.25	100	97.37
	Testing (%)	98.56	99.47	98.87	97.06
Stage 2 PPV	Training (%)	87.76	90.32	100	95.6
	Testing (%)	88.73	92.03	94.47	90.04
Stage 3	Training (%)	75.73	75.61	100	98.26

PPV	Testing (%)	74.64	74.31	89.51	86.78
Stage 4 PPV	Training (%)	85.71	80.74	100	96.24
	Testing (%)	86.39	82.54	89.19	86.83
Overall PPV	Training (%)	86.96	86.48	100	96.87
	Testing (%)	87.08	87.09	93.01	90.18

Table 4.6: The achieved NPV results of integrated computational methods

Integrated model		KLD-LDA	KLD-DLDA	KLD-KNN	KLD-LSSVM
Stage 1 NPV	Training (%)	15.22	13.11	100	100
	Testing (%)	13.43	12.64	81.82	100
Stage 2 NPV	Training (%)	31.25	32.67	100	100
	Testing (%)	32.09	34.78	79.66	100
Stage 3 NPV	Training (%)	58.24	51.79	100	96.2
	Testing (%)	55.07	49.7	82.46	86.27
Stage 4 NPV	Training (%)	65.33	67.8	100	100
	Testing (%)	66.36	70.11	81.32	94.37
Overall NPV	Training (%)	42.51	41.34	100	99.05
	Testing (%)	41.74	41.81	81.32	95.16

Table 4.7: The prediction results of individual and integrated classification models

Classification method		LDA	DLDA	KNN	PCA-LDA	PCA-DLDA	PCA-KNN
Accuracy	Training (%)	52.94	51.11	100	54.75	55.2	100
	Testing (%)	51.45	50.36	88.41	53.62	55.07	91.3
Sensitivity	Training (%)	77.78	76.19	100	80.00	50.00	100
	Testing (%)	75.00	75.00	83.33	66.67	51.67	83.33
Specificity	Training (%)	83.96	83.82	100	89.1	89.57	100
	Testing (%)	82.95	83.71	98.48	88.64	89.39	99.62
PPV	Training (%)	17.07	17.2	100	25.81	18.52	100
	Testing (%)	16.67	17.31	71.43	21.05	15.15	90.91
NPV	Training (%)	98.89	98.76	100	98.95	97.42	100
	Testing (%)	98.65	98.66	99.24	98.32	97.12	99.25

Table 4.8: Ranking of features for primarily biliary cirrhosis stages

Rank	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
1	Ascites	Ascites	Ascites	Ascites
2	Hepato	Edema	Chol	Edema
3	Spiders	Protime	Edema	Hepato
4	Edema	Copper	Protime	Copper
5	BIL	Spiders	ALB	ALB
6	Protime	Hepato	ALK	Protime
7	Trig	BIL	Copper	Spiders
8	Chol	ALB	Trig	Chol
9	Copper	Platelet	Age	BIL
10	AST	Chol	Platelet	Platelet
11	ALB	Trig	Hepato	ALK
12	ALK	AST	Gender	Trig

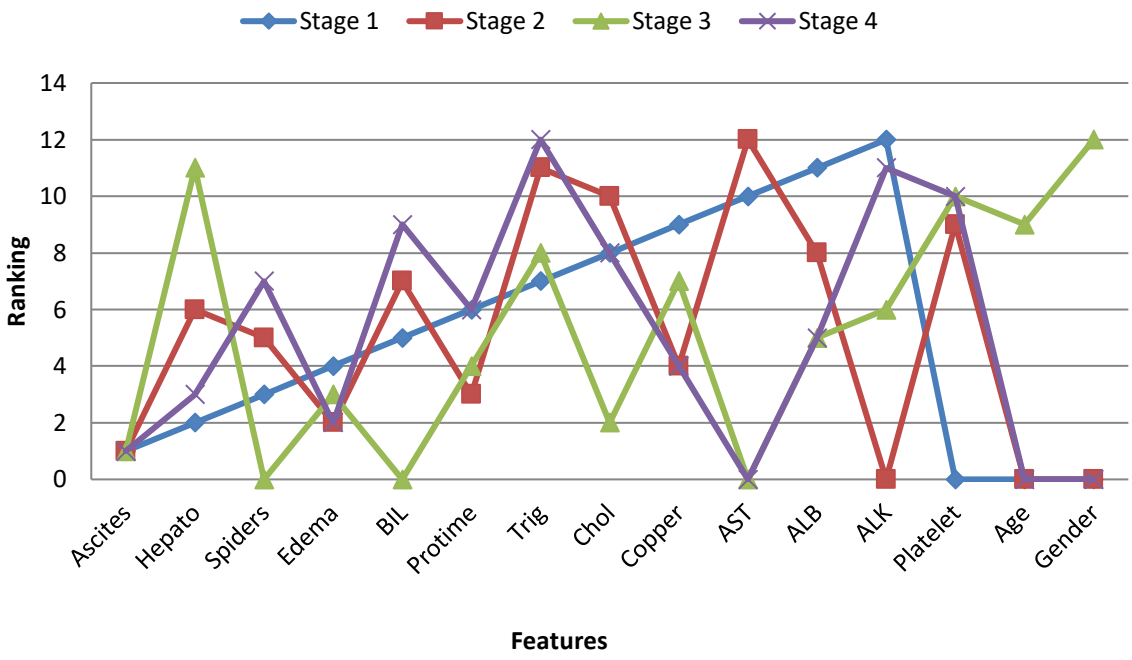

Figure 4.34: Correlation among features with respect to different stages

To select the most efficient feature selection and feature extracted based integrated classification models, obtained results are compared with each other in their category. Table 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7 present the simulation results. It is observed from experiments that KLD-LSSVM in feature selection category and PCA-KNN in feature extraction category outperforms all other classifiers and are selected as best predictive models for PBC stages. KLD based integrated computational methods incorporated features given in table 4.8 for classifying stage 1, 2, 3 and 4. Variation in ranking of features and correlation among features with respect to PBC stages indicated the complex nature of diagnosis process (Figure 4.34). For instance; ascites, hepato, spiders, edema, BIL, protime, trig, chol, copper, AST, ALB and ALK were the key features for stage 1 prediction. Ascites, edema, protime, copper, spiders, hepato, BIL, ALB, platelet, chol, trig and AST were vital of stage 2. Ascites, chol, edema, protime, ALB, ALK, copper, trig, age, platelet, hepato and gender were crucial for stage 3. Ascites, edema, hepato, copper, ALB, protime, spiders, chol, BIL, platelet, ALK and trig were essential for stage 4. Experimental results also showed that exclusion of age, gender and platelet count features did not lessen the prediction accuracy of stage 1. Likewise, leaving out age, gender and alkaline phosphotase in stage 2; removal of spiders, serum bilirubin and aspartate aminotransferase in stage 3; and elimination of age, gender and aspartate aminotransferase in stage 4 showed no decline in identifying PBC stages. Presence of ascites and edema were the most common and influential attributes in all four stages. As the disease

progresses, amount of cooper in urine starts increasing which indicates a sign of being in the end stage of PBC. In summary, the feature selection based study is proved to be an efficient medical support system for predicting PBC stages, finding ranking of features and distinguishing the association of features in priority with different stages. The feature extraction based approach combined advantages of both PCA and KNN such as high classification rates, good generalization, plain structure and efficient problem solving ability through feature extraction. Primary biliary cirrhosis is a general cause of liver cirrhosis globally and will surely keep engaging the novice researchers and physicians in its assessment.

4.4 Conclusions

Liver is an essential organ of human body. Any kind of damage to it can lead to severe medical complications. Physicians have a prime role in the assessment of patients' health but a huge of amount of medical experience is needed to diminish the chances of erroneous diagnosis. Therefore, intelligent computational diagnostic systems have been frequently developed to assist clinicians in evaluating medical records. These systems positively influence the expert decisions by acting as a second opinion. This chapter also put an effort in that direction by proposing the KLD-LSSVM and PCA-KNN based intelligent integrated models for an effective classification of PBC stages. Transformation of features into new space by PCA makes the classifier more efficient. Learning process in KNN is almost zero whose simplest variation is 1NN. Here the value of K is equal to 1. It aims to find the nearest neighbor *n* based on the training dataset. Datasets having large dimensionality is easily handled by single nearest neighbor. The value of K needs to be change in order to condense the noise effect in samples. KNN also solves the problem of over-fitting by searching multiple samples as nearest neighbor. It also has a disadvantage of degradation in accuracy rates with complex dimension datasets. Therefore, dimensionality reduction method was applied in the pre-processing step for boosting classifier performance. Feature ranking by KLD helped in selection of vital attributes needed for input to SVM. SVM has a very high generalization performance and there is no requirement to add a prior knowledge, even when it has very high input space dimension. The main intend of the SVM was to discriminate between members of classes in training data by finding best classification function. 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 is constructed which maximizes the

margin between data sets. Parallel hyper-planes were constructed in order to calculate the margin, one on each side of separating hyperplane. The largest distance to neighboring data points of the classes helped to achieve good separation and if the margin is large then generalization error eventually become less. Thus the support vectors and margins help to find hyperplanes. PBC represents earlier stages of liver illness since complete cirrhosis occurs only in the later stage. It is also known as primary biliary cholangitis and is considered to be of autoimmune in nature because of the presence of autoantibodies. It is a chronic disease where the bile ducts are slowly destroyed within the liver which eventually leads to cirrhosis. This disease is found in every five people out one lakh. It is widely found in women having prevalence rate of ninety percent. Visibility of the disease generally observed between the ages of forty and sixty years and required to be identified early. In summary, it is concluded that the proposed methodologies can also be productively applied to real life health examination datasets, other that PBC, containing a variety of features and multiple decision classes.

The findings of the chapter have been published in the International Journal of Computational Biology and Drug Design, Vol. 10, No. 1, pp. 24-38, 2017, Inderscience Publishers, United Kingdom, DOI: 10.1504/IJCBDD.2017.082807 and in Proceedings of the IEEE: International Conference on Inventive Computation Technologies (ICICT), August 26-27, 2016, Coimbatore, India, Vol. 1, IEEE, DOI: 10.1109/INVENTIVE.2016.7823222, mentioned under the list of publications at the end of the chapter 8.

Chapter 5

An Euclidean Distance Function based Computational Model for Assessing Degree of Liver Damage

This chapter is organized as follows: Section 5.1 introduces the chapter. Section 5.2 outlines the computational methodology applied to assess degree of liver damage. Section 5.3 describes the used dataset, discusses the experimental results and compares the performance of proposed approach with other classification models. Finally, section 5.4 concludes the chapter.

5.1 Introduction

Liver is one of the vital organ of human body. It performs numerous metabolic functions that are essential for living a healthy life. Improper functioning of any of the functions leads to liver disease. Liver disease can be acute (for short time) or chronic (for long time) that can put human life at risk [4]. Severity of the disease may begin from a healthy individual to viral hepatitis infection, to cirrhosis and more seriously to liver cancer. Liver disease is one of the leading cause of mortality worldwide. It is a serious area of concern in the universal set of medicine. Presence of the disease is steadily increasing over the years irrespective of age, sex, region or race. Liver resists early detection, as it functions normally even when partially damaged, makes the disease even more alarming because by then it might have suffered eternal damage. This indicates the early diagnosis of liver disease so that in time treatment can be initiated. Various modes of liver diagnosis are liver biopsy, image scan (CT, MRI, ultrasonography etc.), liver function tests, medical history and physical examination [16], [31], [136]. Liver biopsy is still the gold standard method used to detect and characterize liver disease but has important sample error issues and subjectivity in the interpretation. Furthermore, it is an invasive method which also raises risk of complications if the mode of sampling is not appropriate [17]. Physical examination and medical history do not replace other diagnostic procedures as these indications are normal unless the disorder is severe. It rather complements the diagnostic decision. MRI scan has also become a popular mode of evaluation for liver disease prediction but it involves a long procedure time and high cost. Ultrasonography unable to differentiate between benign and malignant liver lesions and it can not to detect small hepatic lesions as it fails in penetrating air or bone. Computer tomography

uses iodinated contrast material which is restricted in patients with renal insufficiency. These scans are limited to axial planes. Imaging techniques do have presence of structure noise in the images which makes it difficult for medical expert to interpret precisely. Imaging devices generates huge amount of data that increases the chance of error occurrence.

Liver function tests majorly help in examining the liver injury. Key parameters in the test include bilirubin, albumin, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, total proteins, gamma-glutamyl transferase, prothrombin time, triglycerides platelet count and so on. These parameters indicate the severity of liver damage. Out of the aforesaid attributes this study incorporated alkaline phosphatase, alanine aminotransferase, asparate aminotransferase and total bilirubin for experimentation; and aims to build an intelligent computational model for assessing degree of liver damage. The predictions will eventually help the physicians in recommending appropriate amount of dose to the patients during treatment. This chapter deployed linear discriminant analysis, diagonal linear discriminant analysis, quadratic discriminant analysis, diagonal quadratic discriminant analysis, and euclidean distance function based k-nearest neighbor classifiers. Performance of all these intelligent classification algorithms is compared in terms of accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Simulations results showed that euclidean distance based k-nearest neighbor model achieved better prediction outcomes than other classifiers. In addition to higher accuracy rates, the model also attained remarkable sensitivity and specificity which is a challenging task given an uneven variance among attribute values in the dataset. The method gives best decisions in shortest time possible and is considered more accurate as the technique used is data driven. The model filters overflow of information, data and knowledge. The approach provides ability to learn from N number of experiences with respect to some task T and to act appropriately in an uncertain environment for increasing the probability of success. It makes the proposed model self modifying and highly automated which continues to improve over time as it learns with more data.

5.2 Methodology

The first step in methodology is loading of dataset, second is preprocessing of data, third is data partitioning, fourth is classification of samples, and last step is performance evaluation and comparison which eventually decides the best computational model for assessing degree of liver damage. In data preprocessing, each sample is symbolized as a vector of real numbers

before giving input to a classifier and the additional features are also removed. For instance, in selector field a sick class is represented by 1 and a healthy class is indicated with 0. The cross validation method partitions the observations into training and testing data, where M observations are randomly selected as an evaluation set. Intelligent classification algorithms implemented in this chapter are linear discriminant analysis, diagonal linear discriminant analysis, quadratic discriminant analysis, diagonal quadratic discriminant analysis and k-nearest neighbor. The prediction performance is calculated using statistical parameters include accuracy, sensitivity, specificity, positive predictive value and negative predictive value. Experimental results show that euclidean distance metric based k-nearest neighbor approach outperforms all other aforesaid classifiers and is selected as the best computational model. Figure 5.1 presents the activity diagram of selected model build for assessing degree of liver damage. Description of all classification algorithms deployed to identify severity of liver damage is as follows.

K-nearest neighbor works on the concept of similarity measures. It preserves a range of cases and then a new case is classified based on various votes acquired from the neighbors. Its cost of the learning process is nearly zero. The simplest variation of KNN is default 1NN model having K=1, where the goal is to find a sample closest to the nearest neighbor y. The idea of single nearest neighbor approach is quite useful when data samples are large in number. The concept of 1NN can be extended to KNN in a way such that higher value of K reduces the effect of noise in given dataset. Instead of searching for a single sample, more samples are searched as nearest neighbor in order to remove the consequences of over-fitting [113], [114], [141]. KNN imposes problems with the increase in number of dimensions. Its prediction accuracy is degraded with the increase in number of attributes. For maintaining the accuracy levels with the increase of attributes various dimensionality reduction techniques are applied in the pre-processing step of classification. Similarity measure in KNN is derived using variety of distance functions. Distance functions used in determining the case with maximum set of nearest neighbors are euclidean, correlation, cityblock, cosine, manhattan and minkowski out of which euclidean function is selected for the given dataset. Euclidean distance function is represented as mentioned in Eq (5.1).

$$Euclidean\ distance\ function = \sqrt{\sum_{j=1}^{k} (x_j - y_j)^2} \tag{5.1}$$

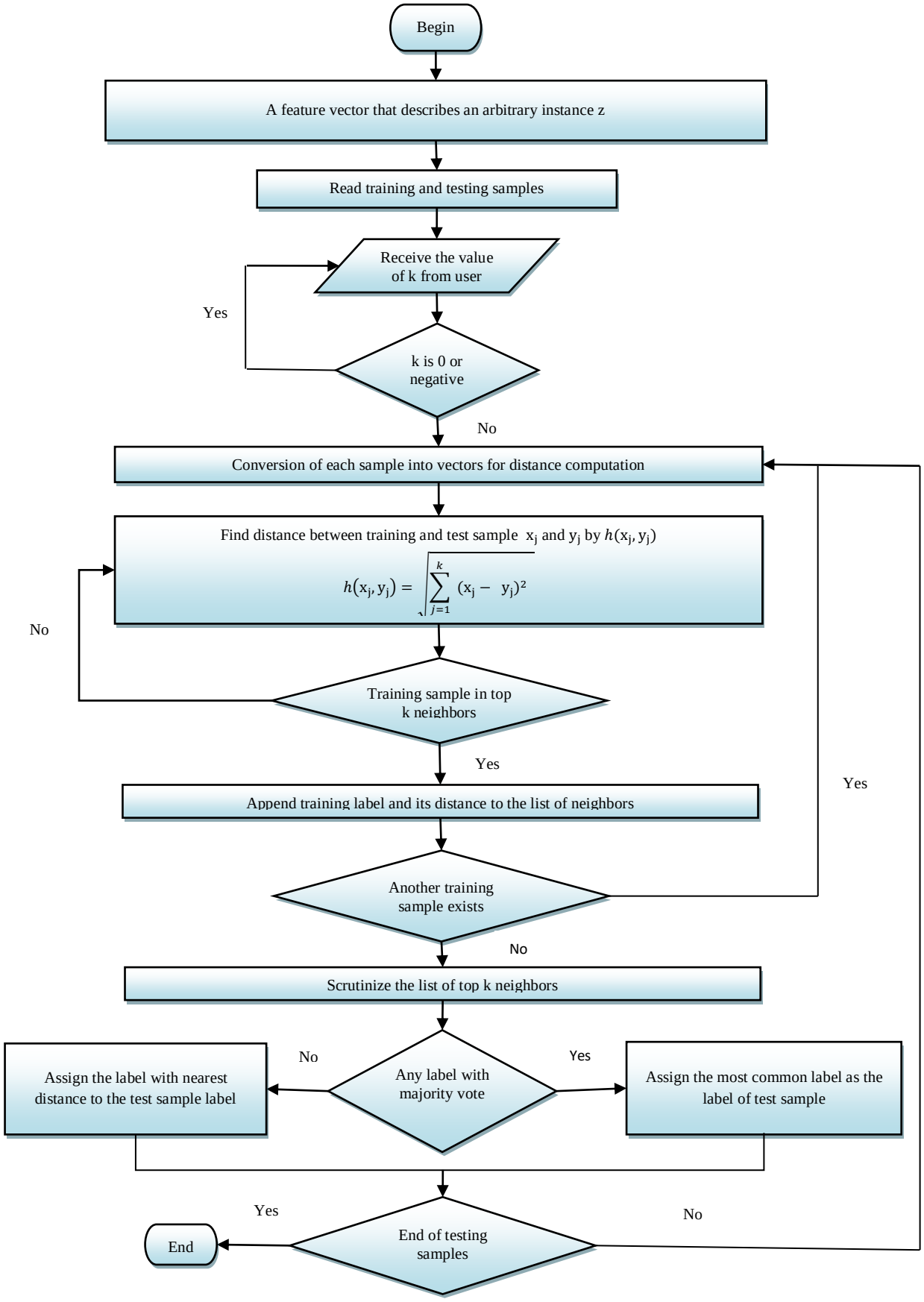

Figure 5.1: Activity diagram of intelligent computational model for assessing degree of liver damage

Linear discriminant analysis works on two variables: dependent and independent. Dependent variable (Y) represents group, and the dataset features that describe a group are independent variables (X). 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. The covariance matrix in LDA remains same for each class but the mean varies [142], [143]. Let us assume that population π_j, with mean vector μ_j and variance-covariance matrix Σ is the probability density function of z. The formula is defined as:

$$f(z|\pi_j) = \frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(z-\mu_j)'\ \Sigma^{-1}(z-\mu_j)\right] \quad (5.2)$$

Population means for each of pooled variance-covariance matrix t and populations μ_j. The linear score function is:

$$a_j^M(Z) = -\frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \mu_j'\Sigma^{-1}z + \log q_j = h_{j0} + \sum_{n=1}^{q} h_{jn}\, z_n + \log q_j \quad (5.3)$$

where, $h_{j0} = -\frac{1}{2}\mu_j'\Sigma^{-1}\mu_j$ and $h_{jn} = \text{nth element of } \mu_j'\Sigma^{-1}$

The left hand expression is similar to regression coefficients h_{jn} and linear regression with intercept term h_{j0}. The linear discriminant function is defined as:

$$h_j^M(z) = -\frac{1}{2}\mu_j'\Sigma^{-1}\mu_j + \mu_j'\Sigma^{-1}z = h_{j0} + \sum_{n=1}^{q} h_{jn}\, z_n \quad (5.4)$$

where $h_{j0} = -\frac{1}{2}\mu_j'\Sigma^{-1}\mu_j$

Discriminant analysis needs estimates of prior probabilities, population means, and variance-covariance matrix. The required parameters in equation (5.4) are estimated as follows:

Prior probabilities:

$$q_j = \Pr(\pi_j);\ j = 1,2,\dots,t \quad (5.5)$$

Population means:

$$\mu_j = E(Z|\pi_j);\ j = 1,2,\dots,t \quad (5.6)$$

Variance-covariance matrix: $\Sigma = \text{var}(Z|\pi_j);\ j = 1,2,\dots,t$ (5.7)

On the other hand, DLDA is the extension of linear discriminant analysis where covariance matrices are assumed equal across groups [107].

Unlike LDA, quadratic discriminant analysis have different covariance matrix and mean for each class. For each class *m*, *m =1, 2, ..., M,* algorithm needs to estimate covariance matrix Σ_m separately. The quadratic discriminant function is defined as:

$$\delta_m(z) = -\frac{1}{2}\log|\Sigma_m| - \frac{1}{2}(z-\mu_m)^C\,\Sigma_m^{-1}(z-\mu_m) + \log\pi_m \qquad (5.8)$$

The above quadratic discriminant function is not really different from linear discriminant function apart from Σ_m, the covariance matrix. The discriminant function used here is a quadratic function and it contain second order terms.

Classification rule:

$$\widehat{G}(z) = \arg\max_m \delta_m(z) \qquad (5.9)$$

In classification rule, find class *m* to maximize the quadratic discriminant function. Quadratic equations in z are boundaries. This algorithm has the ability to fit data more suitably as compare to LDA because of its higher flexibility for the covariance matrix but there is need to estimate more parameters. As QDA have different covariance matrix for each class, the problem arises if there are more classes and not so many sample points [108], [144]. On the other hand, DQDA is the extension of quadratic discriminant analysis where covariance matrices are used in which all off-diagonal elements are set to be zero.

5.3 Results and Discussion

A clinical trial laboratory dataset containing six hundred and six medical records of liver patients obtained from AstraZeneca data records is taken for experimentation. The dataset signifies four doses (A, B, C, D) of a drug. These doses are referring to four different degrees of liver damage. The patient receiving dose A has lowest degree of liver damage and patient receiving dose D has the highest degree of damage. The clinical data is having total nine attributes out of which four are removed in data preprocessing and finally remaining five are incorporated out of which four are input variables and one is target variable. These input features include alkaline phosphatase (ALK), alanine aminotransferase (ALT), asparate aminotransferase (AST) and total bilirubin (TBL). Each sample in the dataset represents a record for any of the four recommended doses. For record, the data has one hundred and fifty two patients with dose A, one hundred and forty eight patients with dose B and C; and one hundred and fifty eight with dose D. To reduce sample biasness, to estimate misclassification probabilities, and to validate the computational models, the given dataset is divided into eighty percent training and twenty percent testing data using holdout cross validation method.

Simulation results obtained from the models are compared using accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) rates that are defined in Eq. (5.10), (5.11), (5.12), (5.13) and (5.14) respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.12}$$

$$PPV = \frac{TP}{TP + FP} \tag{5.13}$$

$$NPV = \frac{TN}{TN + FN} \tag{5.14}$$

where TN indicates true negative (normal people correctly recognized as normal), TP is true positive (diseased people correctly recognized as diseased), FN is false negative (diseased people incorrectly identified as normal), and FP expresses false positive (normal people incorrectly identified as diseased).

Figure 5.2, 5.3, 5.4, 5.5 and 5.6 illustrate the performance comparison among computational models using accuracy, sensitivity, specificity, PPV, and NPV rates respectively. Figure 5.2 depicts that LDA had 28.04% (training) and 28.38% (testing) accuracy; DLDA had 27.42% (training) and 27.06% (testing) accuracy; QDA had 33.4% (training) and 32.51% (testing) accuracy; DQDA had 32.37% (training) and 31.52% (testing) accuracy; and euclidean distance metric based KNN approach had 100% (training) and 92.53% (testing) accuracy. Figure 5.3 shows that LDA had 42.62% (training) and 42.76% (testing) sensitivity; DLDA had 36.36% (training) and 37.5% (testing) sensitivity; QDA had 77.05% (training) and 74.34% (testing) sensitivity; DQDA had 56.2% (training) and 56.58% (testing) sensitivity; and euclidean distance metric based KNN approach had 100% (training) and 94.83% (testing) sensitivity. Figure 5.4 illustrates that LDA had 64.19% (training) and 63.66% (testing) specificity; DLDA had 70.33% (training) and 69.82% (testing) specificity; QDA had 39.67% (training) and 39.65% (testing) specificity; DQDA had 62.91% (training) and 61.23% (testing) specificity; and euclidean distance metric based KNN approach had 100% (training) and 85.65% (testing) specificity. Figure 5.5 shows that LDA had 28.57% (training) and 28.26% (testing) PPV; DLDA had 28.95% (training) and 29.38% (testing)

PPV; QDA had 30.03% (training) and 29.2% (testing) PPV; DQDA had 33.5% (training) and 32.82% (testing) PPV; and euclidean distance metric based KNN approach had 100% (training) and 95.25% (testing) PPV. Figure 5.6 depicts that LDA had 76.9% (training) and 76.86% (testing) NPV; DLDA had 76.88% (training) and 76.94% (testing) NPV; QDA had 83.72% (training) and 82.19% (testing) NPV; DQDA had 81.21% (training) and 80.81% (testing) NPV; and euclidean distance metric based KNN approach had 100% (training) and 84.69% (testing) NPV.

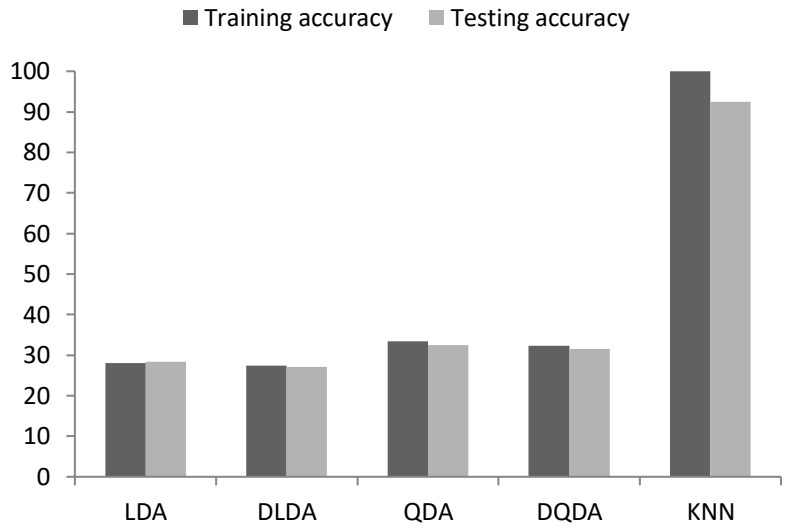

Figure 5.2: The comparative view of obtained accuracy rates

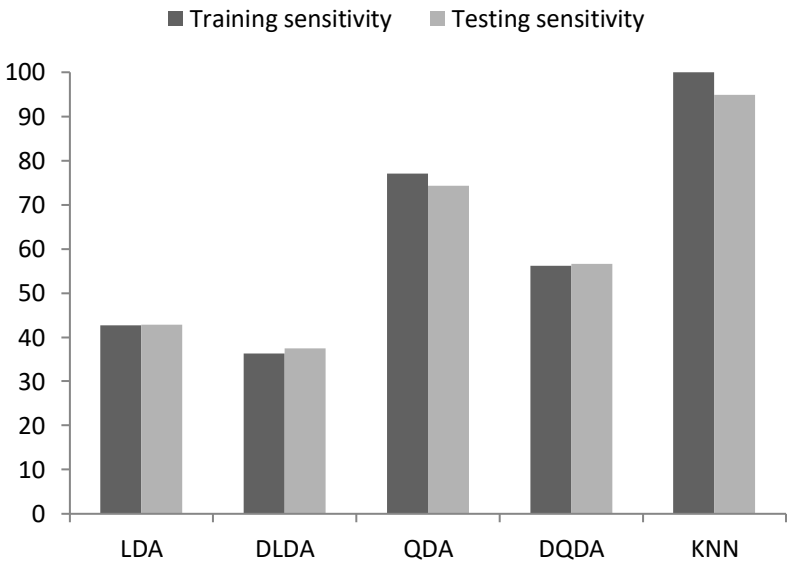

Figure 5.3: The comparative view of obtained sensitivity rates

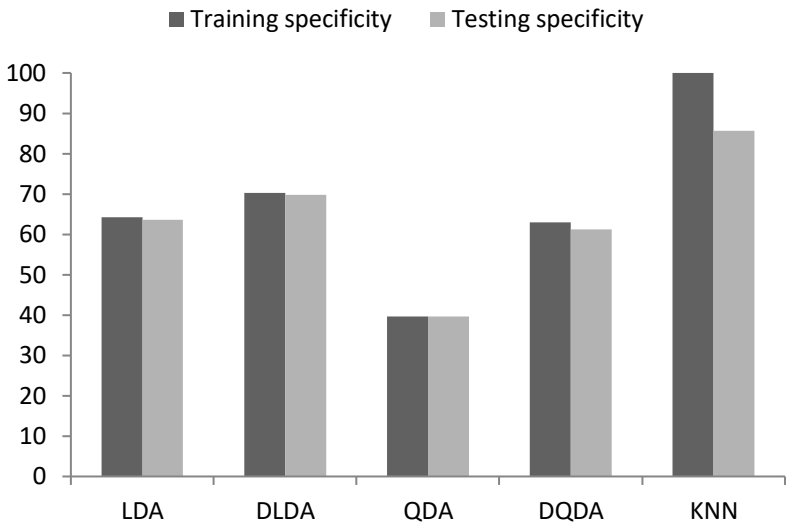

Figure 5.4: The comparative view of obtained specificity rates

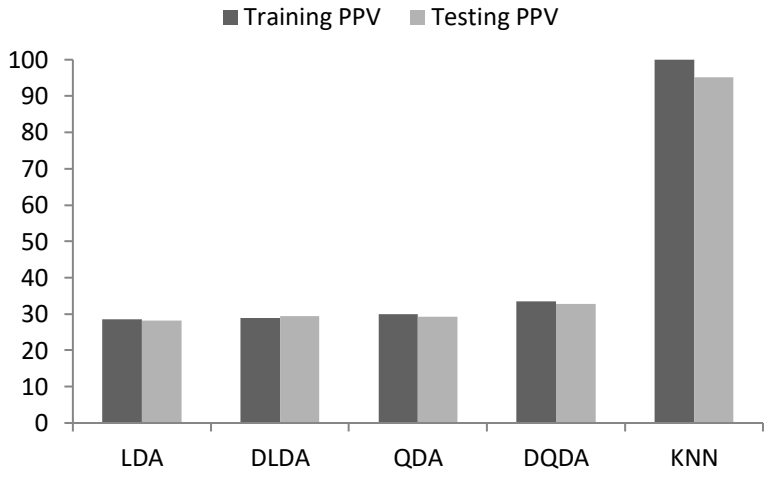

Figure 5.5: The comparative view of obtained positive predictive value rates

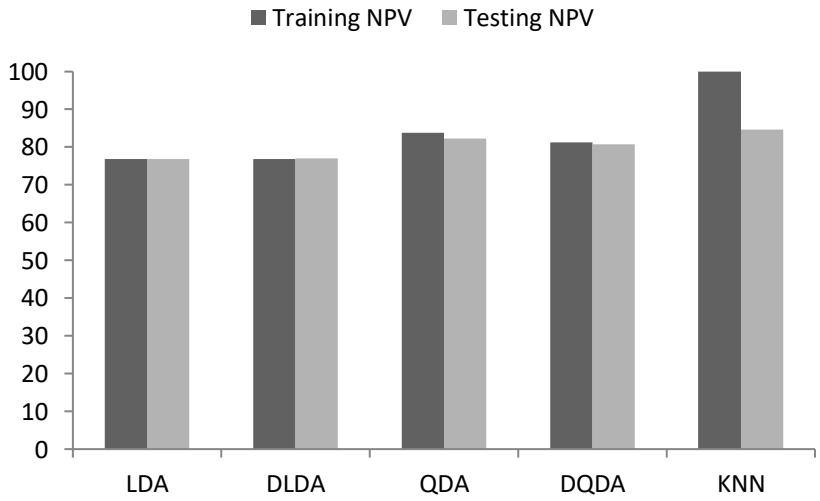

Figure 5.6: The comparative view of obtained negative predictive value rates

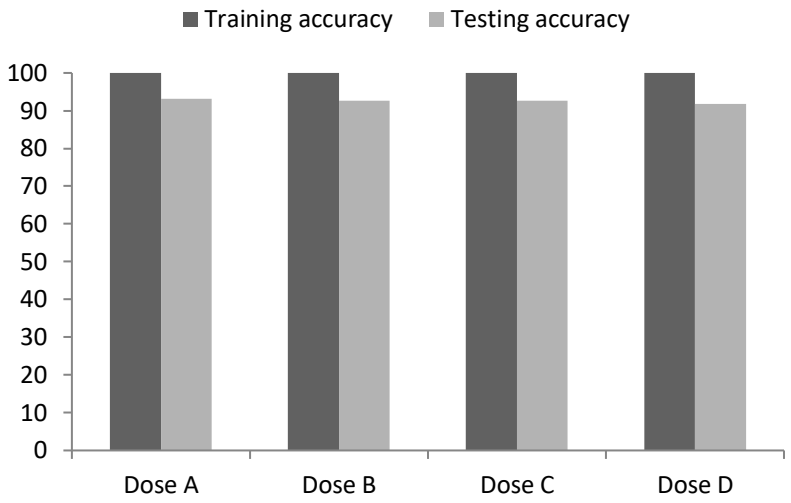

Figure 5.7: The achieved accuracy rates for various degrees of liver damage

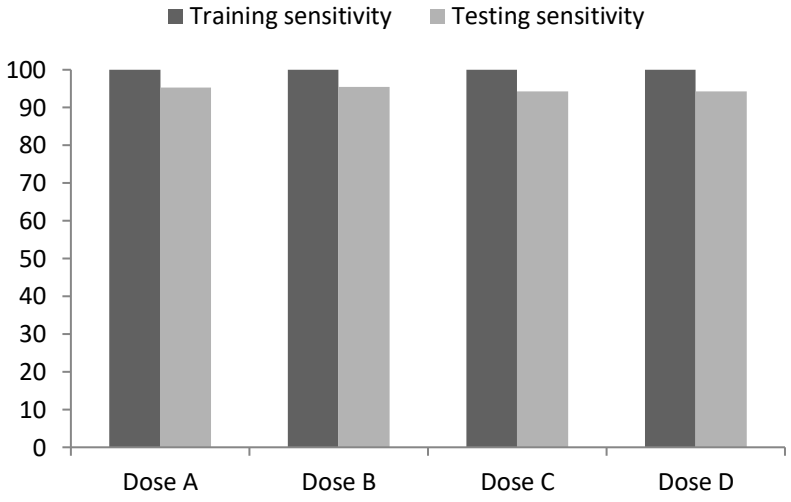

Figure 5.8: The achieved sensitivity rates for various degrees of liver damage

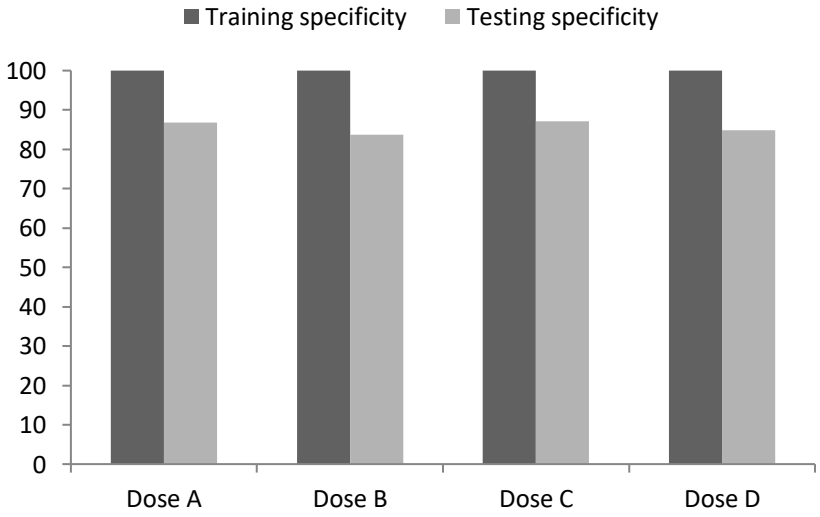

Figure 5.9: The achieved specificity rates for various degrees of liver damage

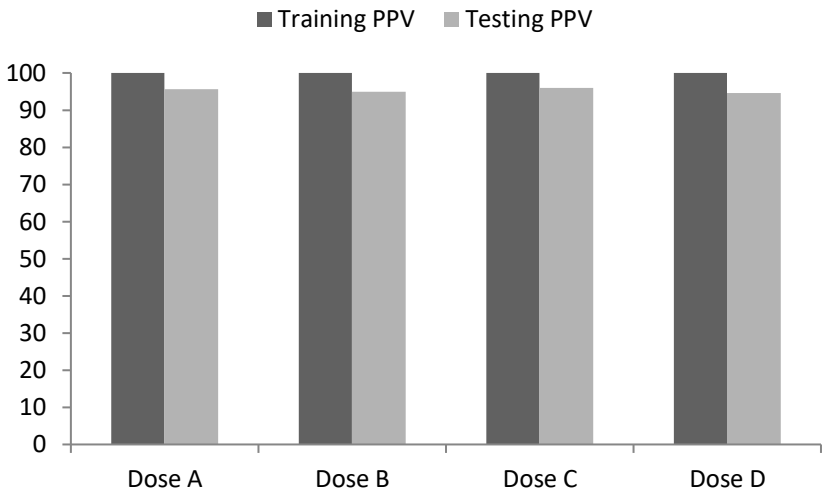

Figure 5.10: The achieved positive predictive value rates for various degrees of liver damage

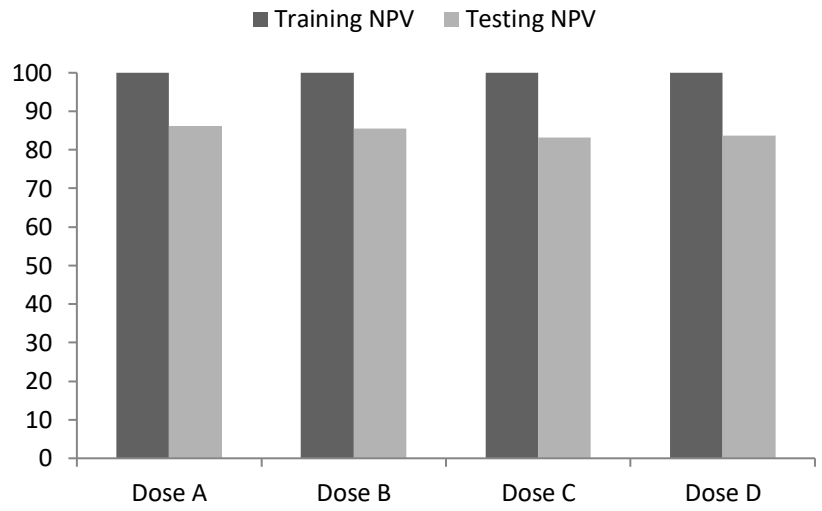

Figure 5.11: The achieved negative predictive value rates for various degrees of liver damage

Experimental results showed that LDA, DLDA, QDA, and DQDA based models had not shown significant diagnostic performance. Only euclidean distance metric based KNN computational model got efficient prediction performance and achieved highest accuracy, sensitivity, specificity, PPV and NPV rates among all (Table 5.1). The dataset signifies four doses (A, B, C, D) of a drug. These doses were referring to four different degrees of liver damage. The patient receiving dose A has lowest degree of liver damage and patient receiving dose D has the highest degree of damage. For all the degrees of damage euclidean distance metric based KNN has produced accuracy rates greater than 90% (Table 5.2). The best diagnostic rates are produced for degree A. Figure 5.7, 5.8, 5.9, 5.10 and 5.11 illustrate the performance comparisons of KNN model for each degree of damage in terms of accuracy,

sensitivity, specificity, PPV, and NPV rates respectively. Figure 5.7 depicts that KNN had 100% (training) and 93.23% (testing) accuracy for dose A; 100% (training) and 92.57% (testing) accuracy for dose B; 100% (training) and 92.57% (testing) accuracy for dose C; and 100% (training) and 91.75% (testing) accuracy for dose D. Figure 5.8 shows that KNN had 100% (training) and 95.37% (testing) sensitivity for dose A; 100% (training) and 95.41% (testing) sensitivity for dose B; 100% (training) and 94.32% (testing) sensitivity for dose C; and 100% (training) and 94.2% (testing) sensitivity for dose D. Figure 5.9 portrays that KNN had 100% (training) and 86.84% (testing) specificity for dose A; 100% (training) and 83.78% (testing) specificity for dose B; 100% (training) and 87.16% (testing) specificity for dose C; and 100% (training) and 84.81% (testing) specificity for dose D. Figure 5.10 illustrates that KNN had 100% (training) and 95.58% (testing) PPV for dose A; 100% (training) and 94.89% (testing) PPV for dose B; 100% (training) and 95.89% (testing) PPV for dose C; and 100% (training) and 94.62% (testing) PPV for dose D. Figure 5.11 presents that KNN had 100% (training) and 86.27% (testing) NPV for dose A; 100% (training) and 85.52% (testing) NPV for dose B; 100% (training) and 83.23% (testing) NPV for dose C; and 100% (training) and 83.75% (testing) NPV for dose D.

Table 5.1: The simulation results of classification models

Classification method		LDA	DLDA	QDA	DQDA	LSSVM	KNN
Accuracy	Training (%)	61.1	62.13	52.15	52.67	76.08	100
	Testing (%)	60.89	61.92	52.14	52.66	75.99	99.83
Sensitivity	Training (%)	78.92	77.11	94.58	96.39	26.51	100
	Testing (%)	78.44	76.65	94.61	96.41	26.35	100
Specificity	Training (%)	53.98	56.14	35.18	35.18	95.9	100
	Testing (%)	53.85	56.01	35.1	35.1	95.91	99.4
PPV	Training (%)	40.68	41.29	36.85	37.3	72.13	100
	Testing (%)	40.56	41.16	36.92	37.35	72.13	99.76
NPV	Training (%)	86.49	85.98	94.19	96.05	76.54	100
	Testing (%)	86.15	85.66	94.19	96.05	76.54	100

Table 5.2: The simulation results for various degrees of liver damage

Degrees of Liver Damage		Dose A	Dose B	Dose C	Dose D
Accuracy	Training (%)	100	100	100	100
	Testing (%)	93.23	92.57	92.57	91.75
Sensitivity	Training (%)	100	100	100	100
	Testing (%)	95.37	95.41	94.32	94.2
Specificity	Training (%)	100	100	100	100
	Testing (%)	86.84	83.78	87.16	84.81
PPV	Training (%)	100	100	100	100
	Testing (%)	95.58	94.89	95.89	94.62
NPV	Training (%)	100	100	100	100
	Testing (%)	86.27	85.52	83.23	83.75

5.4 Conclusions

Human decisions are considered to be subjective as it depends on individual judgment and preference. Human mind does have limitations of recalling crucial details required to solve a problem. Certainly, clinicians play a decisive role in medical diagnosis and treatment. However, deployment of intelligent techniques-based models enhances the prediction efficiency and also facilitates physicians to make sound judgments on the presence of sickness. These models produce fair and consistent decisions based on their learning and reasoning capabilities. It extends support to explore information, to manage compound objects, to distinguish the meaning of information available, to improve clinical decisions rather than replacing the judgments. Intelligent techniques provide computers the ability to learn from N number of experiences with respect to some task T and to act suitably in a vague situation for increasing the likelihood of success. It makes the models self-modifying and highly automated which continues to improve over time as it learns with more data. The euclidean distance function and nearest rule based KNN model had given best decisions in shortest time possible and is considered more efficient as the IT used is data driven. The model filters overflow of information, data and knowledge. This algorithm is widely used when the prior knowledge about the distribution of data is unknown. It assigns weights to contributors of nearest neighbors which help in calculating average. These neighbors are taken from the set of objects having correct classification. This step act as a training set and there is no separate training step required. Euclidean distance was used for finding distant measure. It is observed that the algorithm is supersensitive to local data structure. It can also be used in case of regression where we have to assign property values for object to observe the average of values of k nearest neighbors.

The findings of the chapter have been published in Proceedings of the IEEE: International Conference on Inventive Computation Technologies (ICICT), August 26-27, 2016, Coimbatore, India, Vol. 1, IEEE, DOI: 10.1109/INVENTIVE.2016.7823222, mentioned under the list of publications at the end of the chapter 8.

Chapter 6

An Intelligent Hybrid Approach for Hepatitis Disease Diagnosis: Combining Enhanced K-Means Clustering and Ensemble learning

This chapter is organized as follows: Section 6.1 introduces the chapter. Section 6.2 describes the methodology used to diagnose hepatitis disease. Section 6.3 details the material used, discusses the experimental results and compares the performance of proposed approach with other classification models. Finally, section 6.4 concludes the chapter.

6.1 Introduction

Hepatitis is an injury to liver due to inflammation of liver cells. This condition can act naturally constraining or can advance to liver cancer, cirrhosis or fibrosis. It is an inflammation and damage to hepatocytes caused by various types of viruses. These viruses are categorized as hepatitis A, hepatitis B, hepatitis C, hepatitis D, hepatitis E and hepatitis G that are also called HAV, HBV, HCV, HDV, HEV and HGV respectively [69], [73]. About 90% of kids have been contaminated with hepatitis A before the age of 10 yrs because of poor sanitary conditions and unhygienic practices. 15% of individuals are contaminated with hepatitis A and have repeating signs over a 6 to 9 months time frame. For instance, 22,700 cases are reported annually in United States which is 38% of all the registered cases. 500 million individuals are evaluated to be contaminated with hepatitis B worldwide. 30% of the cases infected with hepatitis B show no signs or side effects. This virus kills around 1.5 million people in a year. Chronic illness happens in 90% of children infected with hepatitis B. 1.8% of adults lives with chronic hepatitis B and has an overall incidence rate of 1.49 per 100000 people in the European regions. 1.25 million people have chronic hepatitis B in United States. By and large the age group from 20 to 49 years suffers with this virus. World health organization estimated that 3% of world's population is contaminated with hepatitis C. 80% of the infected people show no signs or side effects. It is generally caused due to usage of high drugs. 4 million people in United States are infected with this virus. 8% of adults live with chronic hepatitis C which causes 86,000 deaths per year in European regions as per the reports of world health organization. HAV is present in the excrement of infected people and is regularly transmitted through utilization of impure water or sustenance. Certain sex practices can likewise spread HAV. Vast majority of people are making a full recuperation

and staying insusceptible from further HAV contamination as infection in many cases is mild. In any case, HAV virus can likewise be serious and life threatening. HBV is transmitted when blood, semen, or another body liquid of an infected person enters the body of a healthy individual. It can also be transmitted from a contaminated mother to the newborn child. Chance for perpetual contamination is identified with age. Around 90% of infected babies turn out to be constantly suffering, contrasted and just 2–6% of them grown-up. Transmission may likewise happen through sexual contact, sharing needles, syringes, or other medication infusion hardware. HCV is usually transmitted through contact with infective blood and injections. People also get stained by sharing needles or other gear to infuse drugs. Sexual transmission is additionally conceivable, yet it is less common. Majority of people won't know about their contamination since it remains asymptotic. There is no immunization for HCV. HDV infects people who are already suffering from HBV. The double contamination of HDV and HBV bring more serious illness. HBV immunizations give assurance from HDV contamination. HEV is transmitted through utilization of polluted water or nourishment. It is a prime reason for hepatitis break out in developing regions and is also perceived as a vital cause of ailment in developed nations. Protected and effective immunizations, to prevent HEV, have been created but are not generally accessible. HGV additionally named GBV-C is late found virus that takes after HCV. The medical facts about the virus are under scrutiny and its part in creating ailment in people is still unclear. In early stages hepatitis may cause flu like symptoms which include general ill feeling, stomach pain, fever, muscle throbs, loss of craving, vomiting, diarrhea, nausea, yellowing of skin and whitening of eyes. However, some infected people do not have symptoms and they might not know whether they are contaminated or not. Symptoms begin to damage liver as the hepatitis progresses. Chemicals emitted by the liver start to develop in blood which causes jaundice, dark urine, foul breath, bitter taste in mouth, light colored stools and abdominal pain. Risk factors of hepatitis include contact with infected individual, exposure to blood and blood fluids, underlying liver ailment, poor cleanliness, insufficient sanitation, smoking and fatty liver. Liquor, drugs, toxins and other infections are the causes of hepatitis virus in body. It also spreads from person to person through contact of blood and by sharing razors, toothbrushes etc. Hepatitis virus in any of the aforesaid form leads to liver damage.

Automatic diagnosis systems play a vital role in predicting disease based on a complex database containing records of healthy and sick people. Clinical interpretations from a collection of symptoms, risk factors, laboratory tests and other vital examination figures is a

highly demanding task in hepatitis diagnosis. The task becomes even more complex if the existing figures are fuzzy. It also stretches the decision time of physicians despite having a huge experience and if they are novice then it may take years for them to gain substantial expertise. Moreover, the accurate diagnosis is still not guaranteed as humans are prone to errors no matter whatever may the reason be like abundant clinical workload or a poor health. Medical datasets having variety of features can increase the complexity of decision process. The applicability of intelligent systems in liver disease prediction has given a better facility to patients as well as doctors for correct and timely diagnosis. Presently in medicine, there are number of computer-aided expert systems available for assessing liver disease. However, still a lot can be explored in developing an intelligent computing model for diagnosing hepatitis disease as it has overlapping symptoms. If a patient has an erroneous health evaluation due to ambiguity of medical attributes, then it would eventually lead to wrong treatment. Therefore, it is a matter of prime concern to correctly predict hepatitis as this disease is also associated with other autoimmune disorders. Deployment of intelligent techniques seems to be essential in order to interpret the complex structure of the given dataset and to productively assist physicians in diagnosis process. Hence, to interpret multifaceted dataset, to avoid clinical inexperience and to reduce the evaluation time; this chapter proposes an intelligent hybrid approach for hepatitis disease diagnosis by combining enhanced k-means clustering and ensemble learning algorithms. The advantage of deploying ensemble learning is that it constructs a set of hypothesis by using multiple learners to solve a specific problem. The proposed system works with combination of data clustering and ensemble learning performed by enhanced k-means clustering, and improved adaptive boosting, bagged decision tree and J48 decision tree respectively. The chapter also deployed various combinations of classifiers in ensemble learning which include REP tree and ZeroR, J48 and random forest, and adaptive boosting and J48. Performance of all the build systems is compared in terms of accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. Simulation results show that the enhanced k-means clustering and improved ensemble learning with enhanced adaptive boosting, bagged decision tree and J48 decision tree based intelligent hybrid approach achieved better prediction outcomes than other individual and integrated methods. In addition to higher accuracy rates, the model also attained remarkable precision and true positive rates.

6.2 Methodology

The study aims to propose an intelligent hybrid approach combining clustering and ensemble learning algorithms for hepatitis disease diagnosis. The diagnostic approaches implemented in this work are represented as EL1, EL2, KM-EL1, KM-EL2, KM-EL3, KM-EL4 and KM1-EL4. EL1 stands for ensemble learning with REP tree as base learner and ZeroR as meta-classifier. EL2 signifies ensemble learning with J48 decision tree as base learner and random forest as meta-classifier. EL3 indicates ensemble learning with adaptive boosting as base learner and J48 decision tree as meta-classifier where learning model of adaptive boosting is decision stump algorithm. EL4 stands for enhanced ensemble learning with improved adaptive boosting and bagged decision tree as base learners and J48 decision tree as meta-classifier where learning model of adaptive boosting is random forest algorithm. KM stands for k-means clustering and KM1 denotes enhanced k-means clustering. KM-EL1, KM-EL2, KM-EL3 and KM-EL4 indicate integration of k-means clustering with EL1, EL2, EL3 and EL4 respectively. KM1-EL4 symbolizes the integration of enhanced k-means clustering and EL4 which is the best diagnostic approach among all aforesaid methods. Figure 6.1 illustrates the block diagram of proposed intelligent integrated system KM1-EL4 for hepatitis disease diagnosis. Firstly, the hepatitis patient data is taken as input in form of raw instances. The dataset contains fifteen attributes with missing values which are filled using a predefined class. Then the data values are converted from numeric to nominal format for giving input to system. Secondly, enhanced k-means clustering algorithm is deployed to cluster the data. Then, the enhanced ensemble learning approach is used to predict hepatitis cases. Advantage of deploying ensemble learning is that it constructs a set of hypothesis by using multiple learners to solve the given problem. It uses enhanced adaptive boosting and bagged decision tree algorithms as base learners and J48 decision tree as meta-classifier. Owing to enhanced clustering, the ensemble model predicts hepatitis cases efficiently. Description of intelligent clustering and classification techniques used in the proposed diagnostic model KM1-EL4 are as follows.

Clustering organizes data objects into multiple groups for showing internal structure of the data. In a single cluster, data objects have high intra cluster similarity and low inter cluster similarity. This similarity is measured in terms of placing of data points by using a distance measure function. K-means is an exclusive clustering algorithm where all the data objects are divided into different clusters and each data point belongs to a separate cluster. It means once

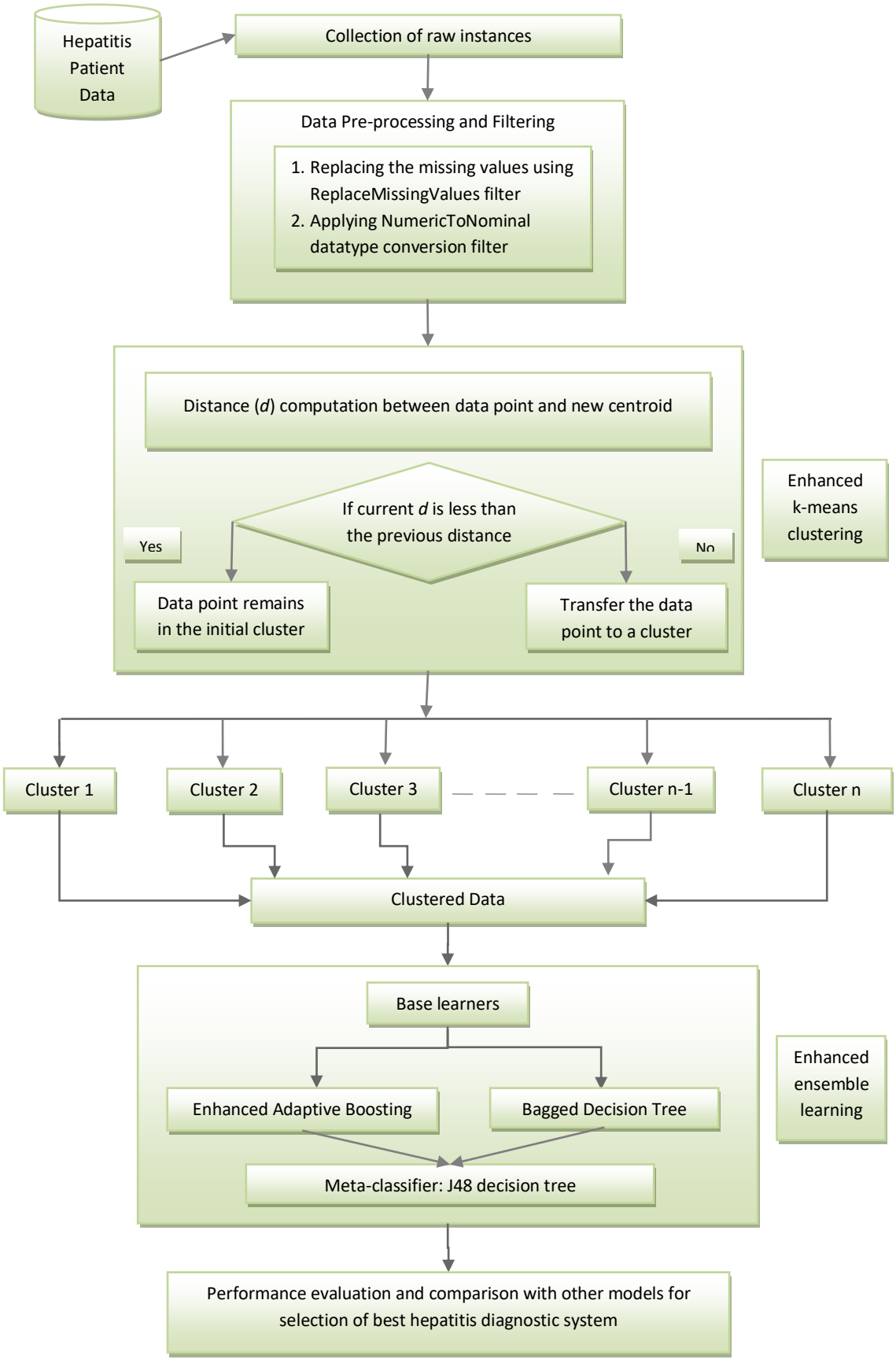

Figure 6.1: Block diagram of proposed intelligent integrated system for hepatitis disease diagnosis

a datum fit in to a cluster then it could not be a part of another cluster. It completely avoids overlapping of data points [145]–[147]. It is unsupervised in nature and used with unlabeled data. Firstly, the number of clusters is selected for giving input to the algorithm. Here, hepatitis patient data is divided into two clusters. K-means begins with randomly selecting initial centroid coordinates or cluster centers for the clusters, then determines the distance of each data object to centroids and then performs grouping of data objects based on minimum distance. An individual data point of hepatitis data would have two calculated distance values. One value is the distance from datum to centroid of first cluster and second value is the distance from datum to centroid of second cluster. The data point is assigned to a cluster center which is closet. Incase if distance values are equal to both centroids then datum is randomly assigned to one of them. In first iteration of algorithm, centroids are randomly selected and then a new centroid is computed by finding average of all data objects that are belonging to the cluster. Reselection of cluster centers continues with each iteration and the distance between data points and new centroids is again calculated for moving them to a more precise location. This process continues until convergence has been reached and finally a bunch of data objects are assigned to each cluster. Convergence is met when the centroids do not move any more. For instance, *k* clusters are assumed apriori for hepatitis patient data. Set of data objects or data points and set of cluster centers or centroids are represented in equation (6.1) and (6.2) respectively. Centroids are recalculated using equation (6.3). The algorithm aims to minimize an objective function which is also known as squared error function given in equation (6.4). In enhanced k-means clustering, storage variables are used in each iteration to retain the label and distance information of each data point to the nearest cluster. This information is used in the next iteration for skipping data object transfer to clusters repeatedly by comparing distance between current datum and new cluster center with the distance of the old center for the same data point. It means, in each iteration, a data point is only transferred to a cluster if its present computed distance is greater than the previous calculated distance stored in the variable, otherwise the datum stays in its cluster which was allocated to it in previous iteration and there is no need to calculate its distances to rest of the cluster centroids. This entire process continues until no changes are detected in centroids. The procedure also saves the computational time as transfer of data points does not take place with each iteration.

$$P = \{p_1, p_2, p_3, \dots\dots\dots p_n\}, \qquad set\ of\ data\ points/data\ objects \tag{6.1}$$

$$V = \{v_1, v_2, v_3, \dots\dots\dots v_c\}, \qquad c\ centroids\ are\ randomly\ selected \tag{6.2}$$

$$v_i = \frac{1}{c_i}\sum_{j=1}^{c_i} p_i, \qquad where\ c_i\ indicates\ number\ of\ data\ objects\ in\ i^{th}cluster \tag{6.3}$$

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(||p_i - v_j||)^2 \tag{6.4}$$

$where, \left|\left|p_i - v_j\right|\right|\ repersents\ the\ distance\ function$

$c_i\ indicates\ number\ of\ data\ objects\ in\ i^{th}cluster$

$c\ is\ total\ number\ of\ centroids$

Ensemble learning constructs a set of hypothesis using multiple learners for solving hepatitis cases. It used improved adaptive boosting and bagged decision tree algorithms as base learners and J48 decision tree as meta-classifier. These learning models are also referred as weak learners. Combination of multiple learning models creates an improved composite classification model. Adaptive boosting in short is called as AdaBoost. It uses training set re-weighting where each training instance uses a weight to find probability of being selected for a training set. It focuses on intricate data points which have been misclassified most by preceding weak classifiers. It builds a strong classifier through linear combination of weak classifiers for reducing bias and variance [148]–[151]. Its working for hepatitis dataset is as mentioned. *P* denotes instance spaces and *Q* indicates set of class labels, assuming *Q* = {-1,+1}. Dataset *D* is used to create training set and an equal weight is allocated to each training instance. A classification model h_t is derived from training samples and its error is calculated using testing. Weights of misclassified training instances are updated to pay more attention in subsequent classification. Incorrectly classified samples lead to increase in their weights and vice-versa. Thus, the weight tends to focus on hard instances. In each round, more focus is given to misclassified instances of previous round. Error rate ϵ_t of classifier h_t is computed by finding sum of weights of each misclassified instance. If the error rate equals or exceeds 0.5, then classifier is terminated and a new training set is generated for deriving a new classifier. This process works in an iterative manner and is repeated *T* times. The final model is derived by weighted majority vote of *T* weak learners. In improved adaptive boosting, firstly random decision forest is used as a learning model instead of decision stump algorithm. Secondly the assigning of weights to misclassified instances is not equal and is divided on the basis of different error rate range. If the obtained error rate is greater than or equal to 0.1 and less than 0.3, then extra weight of 0.1 is added; and if the error rate is greater than or equal to 0.3 and less than 0.5, the extra weight of 0.2 is added. Thus more weight is assigned to

misclassified instances having more error rates which increase their probability of being correctly classified in the next round.

$$D = \{(p_1, q_1), (p_2, q_2), \dots, (p_m, q_m)\};\ where\ p_i \in P\ and\ q_i \in Q \tag{6.5}$$

$//\ p_i\ is\ an\ instance\ drawn\ from\ set\ P\ and\ q_i\ is\ the\ target\ class\ linked\ with\ p_i$

$$D_1(i) = 1/m\,;\quad //\ weight\ distribution\ initialization \tag{6.6}$$

$$For\ t = 1, \dots, T:\quad //\ T\ is\ number\ of\ learning\ rounds \tag{6.7}$$

$$Find\ h_t = arg\ \min_{h_t \in H} \epsilon_t = \sum_{i=1}^{m} D_t\,(i)[\![q_i \neq h_t(p_i)]\!] \tag{6.8}$$

$h_t\ is\ a\ base\ or\ weak\ learner\ which\ is\ trained\ from\ D\ using\ distribution\ D_t\ and$
$\epsilon_t\ is\ the\ weighted\ error\ rate\ of\ h_t$

$If\ \epsilon_t \geq 1/2\,, then\ abort\ loop$

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right);\ //\ weight\ determination\ of\ h_t \tag{6.9}$$

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t q_i h_t(p_i)}{Z_t}\quad //\ Distribution\ updation \tag{6.10}$$

$where\ Z_t\ is\ a\ normalization\ factor$

$$\exp\big(-\alpha_t q_i h_t(p_i)\big)\begin{cases} < 1, & q_i = h_t(p_i) \\ > 1, & q_i \neq h_t(p_i) \end{cases}$$

$//\ weight\ reduction\ of\ correctly\ classified\ samples$

$//\ weight\ increament\ of\ misclassified\ samples$

The final output of the classifier is represented as:

$$H(p) = sign\big(f(p)\big) = sign\left(\sum_{t=1}^{T} \alpha_t h_t\,(p)\right) \tag{6.11}$$

Bootstrap aggregating improves result of final model by considering the majority vote from models trained on bootstrap samples of training data. It works well for overfit models as it decreases the variance without changing bias [152]–[155]. For hepatitis dataset, it draws *B* bootstrap samples of training data through random sampling with replacement, retrains the model on each bootstrap sample and then classifies sample on majority vote. Thus for *B* variations of training set we get *B* particular classifiers. The size of random sample is same as that of training set. J48 decision tree algorithm is used as the classification learning scheme in bagging. For two hepatitis classes live or die, the decision tree algorithm for a training set *D* creates a classifier $H: D \rightarrow \{-1,1\}$. A sequence of classifiers H_m is generated by bagging

method $where\ m = 1, \dots, M$ in respect to modifications of training set. All the learning algorithm based classifiers are coupled to form a final combined classifier whose prediction is given as a weighted combination of individual classifier predictions. This is represented as:

$$H(d_i) = sign\left(\sum_{m=1}^{M} \alpha_m H_m\,(d_i)\right) \tag{6.12}$$

Here classification of instance d_i to either live or die class depends on the majority of classifiers vote where most voted class is predicted. α_m values are found in a way that more accurate models have stronger influence on final decision. The pseudo code of bagging algorithm is as mentioned. Let's assume *m* instances from training set are uniformly sampled with replacement scheme for generating a bootstrap sample. Total T bootstrap samples are generated which are symbolized as $B_1, B_2, \dots, B_T$. For B variations of bootstrap samples we get C classifiers where C_i is built from B_i. The final composite model C^* output is based on the most often predicted class by its sub-classifiers $C_1, C_2, \dots, C_T$.

Input: D, L and T which indicates training set, base learning algorithm and number of bootstrap samples respectively.

$For\ i = 1\ to\ T$:

$D' = D$ $\quad // D'\ is\ bootstrap\ sample\ obtained\ with\ replacement\ scheme\ from\ D$

$C_i = L(D')$ $\quad // Base\ learning\ algorithm\ training\ from\ bootstrap\ sample$

end

Output:

$$C^*(x) = arg \max_{y \in Y} = \sum_{i:C_i(x)=y} 1 \quad // y\ is\ most\ often\ predicted\ label \tag{6.13}$$

J48 algorithm uses a greedy technique to create binary tree structure in a top-down recursive divide-and-conquer manner for classification. It is a simple version of C4.5 decision tree. Based on labelled training data, it uses the normalized information gain (information gain ratio) of attributes for building decison tree. It is a supervised classification technique where the splitting procedure stops once all samples in a subset belong to same class. The tree representation has a nonleaf node, a leaf node and branch which denotes a test on an attribute, a class prediction and an outcome of the test respectively. Best attribute is chosen at each level or node for partitioning the data into target classes and the top most node is known as root node. The algorithm also removes irrelevant attributes in decision tree constuction [153], [156]–[159]. For classification of a test instance, its feature values are tested by following a path from root to a leaf node holding the target class prediction. It works with three

parameters: training instances (*D*) with associated class labels, list of attributes describing each instance and an attribute selection method to decide the spliiting criterion (for deciding the best attribute to distinguish an instance). The selection procedure provides ranking to attributes based on the scores measured. Information gain ratio is the attribute selection measure used in the algorithm which is described as follows. *D* repersents hepatitis training data partition with class-labelled rows. For *n* distinct values in a class label, *n* distinct classes (S_i) are defined where $i = 1, \dots, n$. $S_{i,D}$ symbolizes the set of tuples of S_i in *D*. Then $|D|$ and $|S_{i,D}|$ indicate total number of instances in *D* and $S_{i,D}$. For calculating the gain ratio (IGR), first the information gain (IG) is to be calculated. *I* denotes the information need to classify an instance in *D*.

$$I(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{6.14}$$

$$I_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \tag{6.15}$$

$$IG(A) = I(D) - I_A(D) \tag{6.16}$$

$$SplitI_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{6.17}$$

$$IGR(A) = \frac{IG(A)}{SplitI_A(D)} \quad // \textit{attribute with maximum IGR is selected as splitting attribute}$$

where p_i is the probability of an instance belonging to class S_i in training data. It is computed by $|S_{i,D}|/|D|$. $I(D)$ is also called as entropy of *D* and it depends on the proportion of instances of each class. $I_A(D)$ is needed to classify an instance from data *D* based on the divisioning by *A*. Log function with base 2 indicates the encoding of information in bits. $\frac{|D_j|}{|D|}$ is the weight of j^{th} partition. *A* is a sub atrribute having *v* distinct values on which the tuples partition in *D* takes place. D_j is number of instances of *D* in class S_i. At node *N*, attribute *A* has been selected as splitting attribute based on highest information gain. When the training data *D* is splitted in to *v* divisions corrospoding to the *v* outcomes of a test on attribute *A*, the information ($SplitI_A(D)$) is attained. For each result, number of instances having that result with respect to total number of instances in *D* is considered where as in IG, the information is measured with respect to achieved classification on same partitioning instances.

6.3 Results and Discussion

Hepatitis disease database obtained from University of California repository of machine learning databases is used for experimentation. The dataset characteristic is multivariate and it includes nineteen attributes, two classes, and one hundred and fifty five instances. Attributes contain information about age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alkaline phosphotase, aspartate aminotransferase, albumin, protime and histology. Dataset contains two target classes: live and die. Each instance in the dataset represents information of a single male or female. K-fold cross validation method with 10 as value of k is used to validate the proposed hepatitis diagnostic system. It reduces the biasness associated with instances through random sampling. K mutually exclusive and equal size subsets are created where the model is trained and tested k times. For each case, one validation subset is tested on the classification model which is trained using k-1 subsets. This process eventually generates k different test results for each training-test configuration where the average of all results provides the final testing accuracy. Obtained results are compared using accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. Principally the output results of a classification model are produced in form of TP, TN, FP and FN; and then the aforesaid parameters are calculated using these values. TP indicates true positive (diseased people correctly recognized as diseased), TN is true negative (normal people correctly recognized as normal), FN is false negative (diseased people incorrectly identified as normal), and FP expresses false positive (normal people incorrectly identified as diseased). Accuracy is the ability to distinguish target classes correctly. It is calculated using the ratio of sum of all TP and TN to sum of all TP, TN, FP and FN. True positive rate is also known as sensitivity or recall which measures the proportion of instances that are correctly classified as class A, among all truly class A instances. It is computed using the ratio of TP to sum of TP and FN. Precision is also known as positive predictive value which measures the proportion of instances that truly belong to class A, among all classified class A instances. It is calculated using ratio of TP to sum of TP and FP. F-measure is also known as F-score which computes performance of a model for positive class. It is calculated using the ratio of multiplication of both precision and recall with 2 to sum of precision and recall. Kappa statistic computes the agreement of prediction with true class. Agreement is scaled between 0.0 and 1.0 where the later value signifies complete agreement. Mean absolute error is an average of absolute errors which is not squared before averaging and it is used to quantify the closeness of predictions to

the eventual outcomes. Unlike MAE, root mean squared error squares the difference between predictions and eventual outcomes before averaging absolute errors in order to assign more weight to large errors. Description of attributes and their value ranges obtained from patient are given in Table 6.1.

Table 6.1: The attributes of hepatitis disease database

Attribute	Description	Values of attribute
Age	Age of the patient	10, 20, 30, 40, 50, 60, 70, 80
Sex	Gender of the patient	Male, Female
Steroid	An organic compound	Yes, No
Antivirals	Drugs used for treating viral infections	Yes, No
Fatigue	Extreme tiredness resulting from illness	Yes, No
Malaise	A general feeling of discomfort	Yes, No
Anorexia	An eating disorder	Yes, No
Liver big	Hepatomegaly	Yes, No
Liver firm	Liver becomes hard	Yes, No
Spleen palpable	Enlargement of the spleen	Yes, No
Spiders	Blood vessel malformations in the skin	Yes, No
Ascites	Presence of ascites	Yes, No
Varices	An abnormally dilated vessel	Yes, No
BIL	Serum bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
ALK	Alkaline phosphatase	33, 80, 120, 160, 200, 250
SGOT	Aspartate aminotransferase	13, 100, 200, 300, 400, 500
ALB	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	Standardised blood clotting time	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	Histological condition	Yes, No

The intelligent diagnostic approaches build and applied for hepatitis disease are represented as EL1, EL2, KM-EL1, KM-EL2, KM-EL3, KM-EL4 and KM1-EL4. EL1 stands for ensemble learning with REP tree algorithm as base learner and ZeroR as meta-classifier. EL2 signifies ensemble learning with J48 as base learner and random decision forest as meta-classifier. EL3 indicates ensemble learning with adaptive boosting as base learner and J48 as meta-classifier where learning model of adaptive boosting is decision stump algorithm. EL4 stands for ensemble learning with improved adaptive boosting and bagging decision tree as base learners and J48 as meta-classifier where learning model of adaptive boosting is random decision forest. KM-EL1, KM-EL2, KM-EL3 and KM-EL4 indicate integration of k-means clustering with EL1, EL2, EL3 and EL4 respectively. KM1-EL4 symbolizes the integration of enhanced k-means clustering and EL4. Figure 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10 and 6.11 illustrate the performance comparison among the hepatitis diagnostic models using accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error, relative absolute error rates respectively.

Figure 6.2 depicts that EL1 had 72.26% accuracy, KM-EL1 had 62.58% accuracy, EL2 had 96.13% accuracy, KM-EL2 had 96.13% accuracy, KM-EL3 had 98.06% accuracy, KM-EL4 had 98.71% accuracy and KM1-EL4 had 99.35% accuracy. Figure 6.3 shows that EL1 had 72.3% true positive rate, KM-EL1 had 62.6% true positive rate, EL2 had 96.1% true positive rate, KM-EL2 had 96.1% true positive rate, KM-EL3 had 98.1% true positive rate, KM-EL4 had 98.7% true positive rate and KM1-EL4 had 99.4% true positive rate. Figure 6.4 portrays that EL1 had 52.2% precision, KM-EL1 had 39.2% precision, EL2 had 96.6% precision, KM-EL2 had 96.6% precision, KM-EL3 had 98.2% precision, KM-EL4 had 98.8% precision and KM1-EL4 had 99.4% precision. Figure 6.5 describes that EL1 had 60.6% f-measure, KM-EL1 had 48.2% f-measure, EL2 had 96.2% f-measure, KM-EL2 had 96.2% f-measure, KM-EL3 had 98.1% f-measure, KM-EL4 had 98.7% f-measure and KM1-EL4 had 99.4% f-measure. Figure 6.6 represents that EL1 had 0% kappa statistic, KM-EL1 had 0% kappa statistic, EL2 had 90.74% kappa statistic, KM-EL2 had 90.74% kappa statistic, KM-EL3 had 95.27% kappa statistic, KM-EL4 had 96.83% kappa statistic and KM1-EL4 had 98.63% kappa statistic. Figure 6.7 depicts that EL1 had 40.24% mean absolute error, KM-EL1 had 46.89% mean absolute error, EL2 had 5.18% mean absolute error, KM-EL2 had 5.39% mean absolute error, KM-EL3 had 2.65% mean absolute error, KM-EL4 had 1.29% mean absolute error and KM1-EL4 had 0.79% mean absolute error. Figure 6.8 presents that EL1 had 44.79% root mean squared error, KM-EL1 had 48.4% root mean squared error, EL2 had 18.94% root mean squared error, KM-EL2 had 18.67% root mean squared error, KM-EL3 had 13.8% root mean squared error, KM-EL4 had 11.36% root mean squared error and KM1-EL4 had 8.04% root mean squared error.

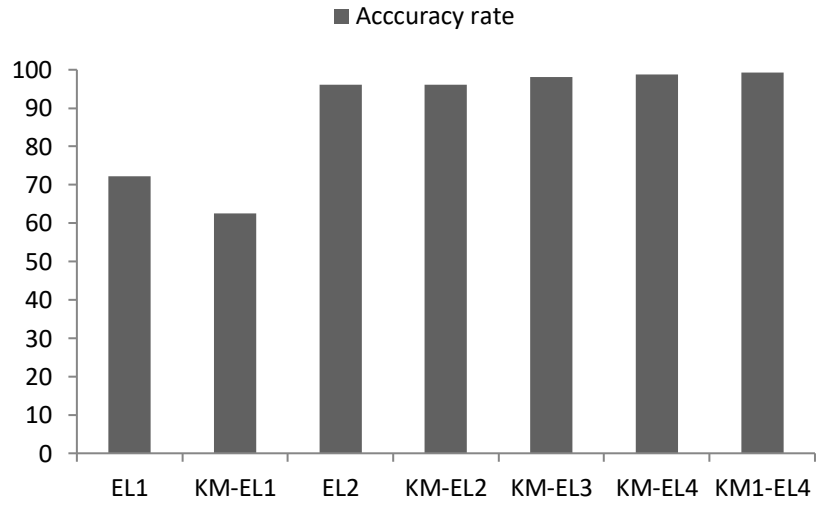

Figure 6.2: The comparative view of obtained accuracy rates

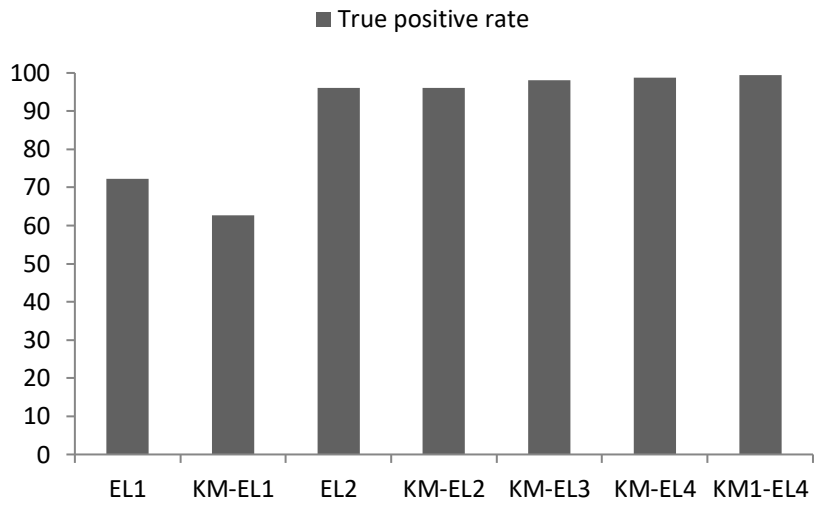

Figure 6.3: The comparative view of obtained true positive rates

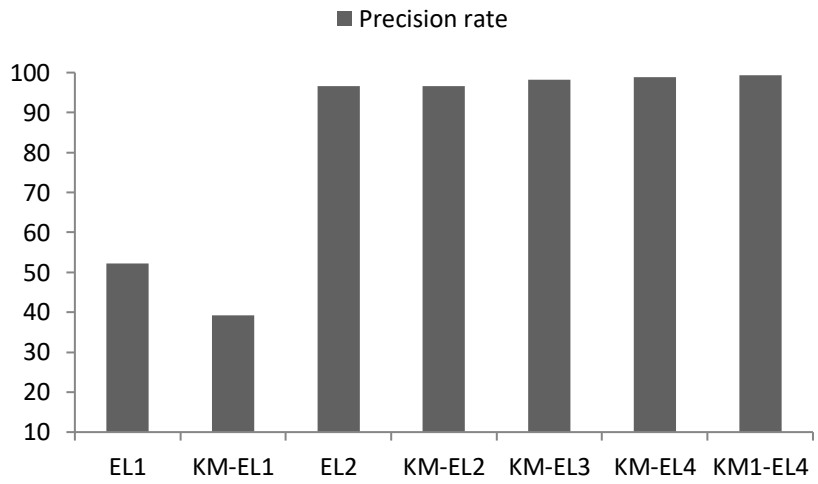

Figure 6.4: The comparative view of obtained precision rates

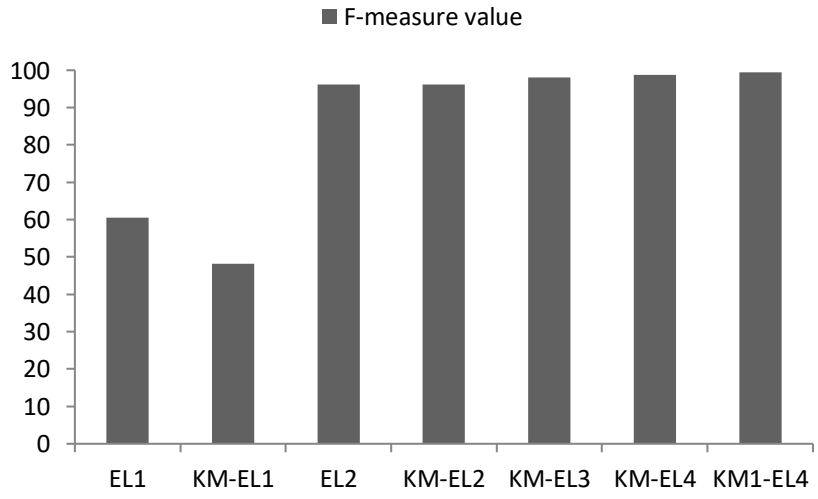

Figure 6.5: The comparative view of obtained F-measure rates

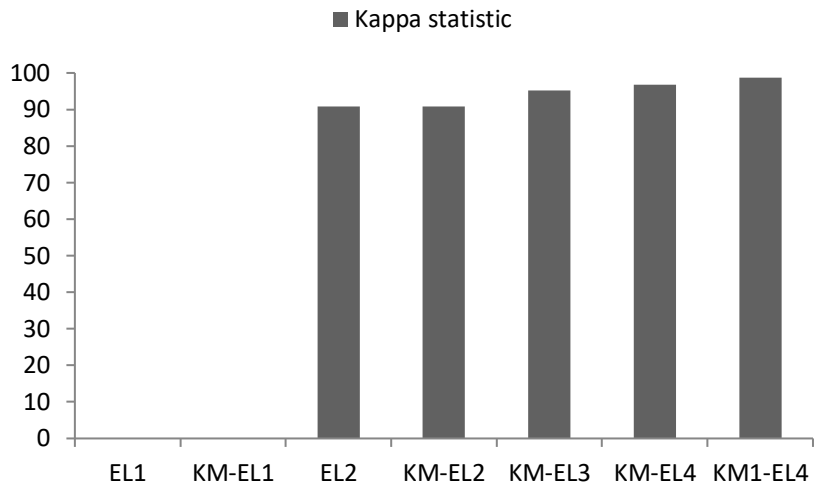

Figure 6.6: The comparative view of obtained kappa statistic

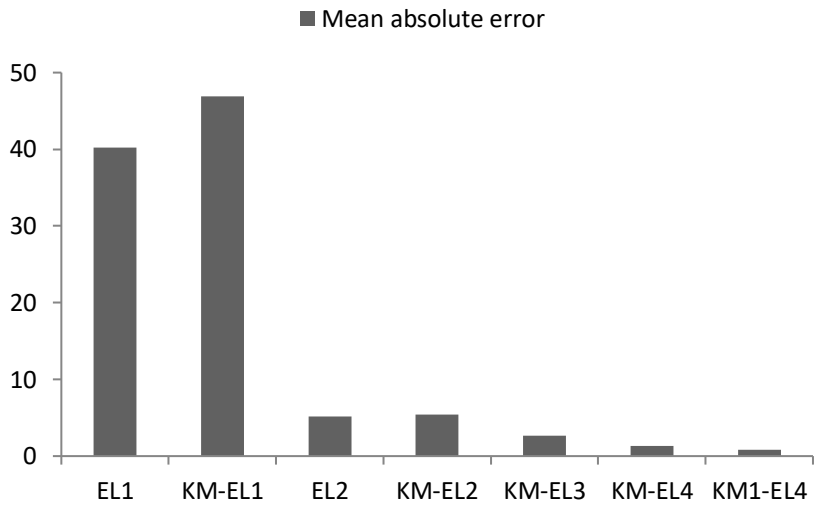

Figure 6.7: The comparative view of obtained mean absolute error rates

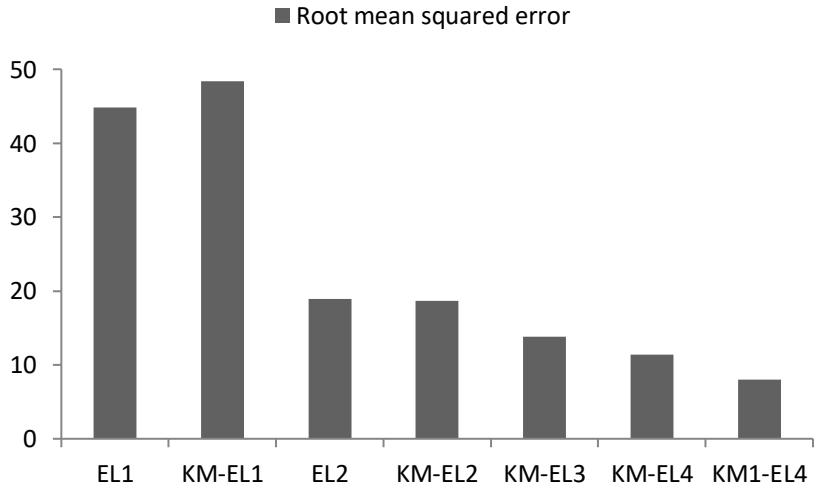

Figure 6.8: The comparative view of obtained root mean squared error rates

To select the most efficient hepatitis diagnosis system, obtained results (accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error rates) of all build models are compared (Table 6.2). Prediction results of KM1-EL4 are also compared to other hepatitis classification methods mentioned in the literature. L. Ozyilmaz and T. Yildirim [27] stated that CSFNN, C4.5, NB, BNND and BNNF achieved accuracy rates of 90.0%, 83.6%, 87.8%, 90.0% and 88.7% respectively. W. Dich, et al. [160] mentioned that weighted 9NN, 18NN and 15NN attained 92.9%, 90.2% and 89.0% respectively. W. Duch, et al. [161] shows that FSM without rotations, RBF and MLP+BP obtained 88.4%, 79.0% and 77.4% respectively. B. Ster and A. Dobnikar [162] cited that LDA, NB and Semi-NB, QDA, ASR, fisher discriminant analysis, LVQ, CART, MLP with BP, and ASI had 86.4%, 86.3%, 85.8%, 85.0%, 84.5%, 83.2%, 82.7%, 82.1%, and 82.0% respectively. K. Polat and S. Gunes [163] stated 92.5% accuracy using FS-AIRS with fuzzy res.; K. Polat, et al. [96] declared 94.1% using FS-fuzzy-AIRS; E. Dogantekin, et al. [73] obtained 94.1% using LDA-ANFIS; M.S. Bascil and F. Temurtas [164] attained 91.8% using MLNN (MLP) + LM; K.C. Tan, et al. [165] achieved 92.4% using CORE; D. Calisir and E. Dogantekin [69] stated 95.0% using PCA-LSSVM; Sartakhti, J. S. et al. [41] stated 96.2% using SVM-SA; H. Kahramanli and N. Allahverdi [93] mentioned 96.78% using an artificial immune system based approach; E. Mezyk and O. Unold [80] showed 93.87 using an artificial immune system with fuzzy partition learning; M. H. Zangooei, et al. [166] obtained 98.52% using support vector regression and a multi-objective evolutionary hybridization; B. Naik, et al. [167] achieved 76.294% using a harmony search based functional link higher order ANN; S. Kulluk, et al. [168] attained 93% using cost-sensitive meta-learning classifier; Q. Hou, et al. [169] stated 91.31% using grouping method based support vector machine; Y. Hayashi and K. Fukunaga [170] mentioned 83.24% using recursive rule extraction algorithm; M. Aldape-Perez [37] showed 85.16% using an associative memory based classifier; S. Ansari, et al. [35] obtained 92% using artificial neural networks; and T. Kanik [171] achieved 94% using rough set approach. Experimental results showed that EL1 and KM-EL1 based diagnostic models have not shown significant performance. Although, EL2, KM-EL2, KM-EL3 and KM-EL4 attained improved accuracy rates then the aforesaid methods but KM1-EL4 achieved highest among all and is selected as the best intelligent integrated approach for hepatitis disease diagnosis. The integrated model combines advantages of k-means clustering, adaptive boosting, bagged decision tree and J48 algorithms such as high classification rates, good generalization, plain structure and efficient problem solving ability. Achieved classification accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute

error and root mean squared error rates of the proposed model are 99.35%, 99.4%, 99.4%, 99.4%, 98.63%, 0.79%, and 8.04% respectively. Generally, clinicians play the prime role in final judgment on patient's health condition but carrying out a resourceful diagnosis is an intricate job that requires enormous medical experience. Certainly, these computationally intelligent diagnostic systems cannot replace physicians' role but may positively assist them in examining medical records by acting as a second opinion. This chapter is also an effort in that direction which proposed an enhanced k-means clustering and ensemble learning based intelligent model for the efficient hepatitis prediction.

Table 6.2: The simulation results of hepatitis diagnostic models

Classification model	EL1	KM-EL1	EL2	KM-EL2	KM-EL3	KM-EL4	KM1-EL4
Accuracy	72.26%	62.58%	96.13%	96.13%	98.06%	98.71%	99.35%
TPR	72.3%	62.6%	96.1%	96.1%	98.1%	98.7%	99.4%
Precision	52.2%	39.2%	96.6%	96.6%	98.2%	98.8%	99.4%
F-measure	60.6%	48.2%	96.2%	96.2%	98.1%	98.7%	99.4%
Kappa statistic	0.0%	0.0%	90.74%	90.74%	95.27%	96.83%	98.63%
MAE	40.24%	46.89%	5.18%	5.39%	2.65%	1.29%	0.79%
RMSE	44.79%	48.4%	18.94%	18.67%	13.8%	11.36%	8.04%

6.4 Conclusions

Physicians have a prime role in assessment of patient health but a huge of amount of medical experience is needed to diminish the chances of erroneous diagnosis. Similarly, evaluation of hepatitis disease is also an intricate task. Deployment of intelligent techniques has contributed a major transformation in the field of information retrieval, and the medical domain has also been widely affected by this renovation. As a part of constant efforts for making hepatitis diagnosis process well-organized and proficient, this chapter built an enhanced k-means clustering and ensemble learning algorithms based hepatitis diagnosis system having an inclusive analytic structure which boosts the prediction performance. Advantage of deploying enhanced ensemble learning is that it constructs a set of hypothesis by using multiple learners to solve the given problem. The intelligent integrated approach showed capability of improving complex medical decisions through clustered data. The prediction was carried out using a data of one hundred and fifty five cases of hepatitis patients. Experimental results confirmed the superiority of proposed approach to other diagnostic models implemented in the chapter and mentioned in literature as well. Mean absolute error and root mean squared error rates were also very small. Thousands people lose

their lives because of erroneous evaluation and inappropriate treatment of hepatitis disease as the medical cases are still largely influenced by the subjectivity of clinicians. The proposed hepatitis diagnostic system can be applied as a liver specialist assistant or as a model to train novice medical students. The system will also help physicians in evaluating complex cases that are otherwise hard to perceive.

Chapter 7

A New Intelligent Model Based on Enhanced Hierarchical Clustering and Random Decision Forest for Classifying Hepatobiliary Disorders

This chapter is organized as follows: Section 7.1 introduces the chapter. Section 7.2 describes the methodologies built to classify alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. Section 7.3 details the material used, discusses the experimental results and compares the prediction performance of proposed approach with other classification models developed in the chapter and mentioned in literature. Finally, section 7.4 concludes the chapter.

7.1 Introduction

Alcoholic liver disease is an injury caused by high alcohol consumption. It starts occurring after an edge measurement of liquor intake is expended. People consume very large quantity of alcohol shows symptoms of liver injury. Alcoholic liver disease is primarily categorized into three stages: fatty liver, alcoholic hepatitis and fibrosis or cirrhosis. The suspicion that the disease dependably advances directly from fatty liver, to alcoholic hepatitis and at later to cirrhosis is not right. Abnormal accretion of fat can occur in parenchymal cells within hours of alcohol drinking in fatty liver disease. Fat deposits collect in central part of the liver as small droplets or large droplets. Thirty three percent of perpetual substantial liquor users fall ill with alcoholic hepatitis and remain asymptotic. Key indications of alcoholic hepatitis incorporate liver cells necrosis, changes in fat and perivenular provocative penetrates. Mallory bodies and eosinophilic accumulations of middle of filaments in the cytoplasm are also found in several patients [172], [173]. Alcoholic fibrosis initially starts in a zone which is pericentral and then advances if damage proceeds. Proceeding with fibrosis and necrosis brings the movement from a miniaturized scale to a macronodular design. This movement is joined by a decrease in fatty liver disease in end-stage. Liver cirrhosis is a condition where the damage is irreversible. Its primary causes are extreme liquor utilization, viral hepatitis B and C, and fatty liver infection. Hepatitis B and C together is said to be a main cause. People with cirrhosis may create jaundice, itching and outrageous tiredness. At a point when liver tissue is destroyed and supplanted by scar tissue the condition becomes serious, as it can begin hindering the stream of blood through liver. Cirrhosis is a dynamic

infection, growing gradually over numerous years, until in long run it stops liver capacity [174], [175]. When it is mild, liver can make repairs and keep working properly but when it is progressed, more scar tissue shapes in liver and then the harm is not repairable. Medications of cirrhosis are aimed for halting or deferring the sickness movement, minimizing liver cell harm and decreasing difficulties. Cirrhosis due to viral hepatitis is treated with antiviral medications to lessen liver cell damage. In any case, a low salt eating routine is additionally vital to treat those with ascites. Drug therapy enhances modified mental capacity connected with cirrhosis. Primary hepatoma is a perilous tumor made out of cells that look like hepatocytes. It is ordinarily attached with cirrhosis and is currently the third major reason for liver cancer worldwide. It especially found in patients with chronic hepatitis B and C. Its prediction can be troublesome and frequently requires utilization of at least one imaging modality. This type of tumors ought to be detected when it is around 2 cm in size. In any case, primary hepatoma is often analyzed later because of the absence of pathognomonic side effects. Therefore, numerous patients with the disease are not treatable when initially identified. The survival period after identification is roughly 6 to 20 months. Extensive tumor measure, poor functional status, nodal metastases and vascular attack are all connected with a poor result. Patients with primary hepatoma have no manifestations other than those identified with chronic liver ailment. Suspicion for the disease ought to be increased in patients with early compensated cirrhosis that creates decompensation. For example: encephalopathy, jaundice, variceal bleeding or ascites. These difficulties are regularly associated with augmentation of tumor into portal or hepatic veins or arteriovenous shunting prompted by the tumor. A few patients may have mild to direct upper stomach suffering, weight reduction, early satiety, or clear mass in the upper belly. These symptoms often indicate an advanced lesion [176], [177] . Cholelithiasis or gallstones signifies as one of the most widely recognized surgical issue around the world and is particularly common in western nations. There are three varieties of gallstones. The first and most regular type is cholesterol stones which represent around 75% of cases. Typically, a delicate balance exists between levels of cholesterol, phospholipids and bile acids. When this balance is upset there is predisposition for the expansion of lithogenic bile and the subsequent development of cholesterol-sort gallstones. Second kind of gallstone is of pigmented mixture. Pigmented stones emerge from crystallization of calcium bilirubinate and happen in two sorts: black and brown. Representing around 15-20% of every single biliary stone, black stones have a tendency to occur in disease related with expanded red platelet destruction and anomalous digestion system of hemoglobin. On other side, brown stones are associated with infected bile and are frequently

found outside of gallbladder. Last kind of gallstone encountered is of mixed variety containing a mixture of pigment and cholesterol. Its most widely recognized side effect is biliary colic which is a serious pain in the epigastrium that endures 1-5 hours and regularly arouses patient from sleep. Nausea, with or without vomiting and flatulence may also be present as symptoms [178], [179].

Computer-aided medical diagnostic systems have been widely practiced in hospitals and are comprehensively assisting physicians in analyzing patient therapeutic history. Large data centers are created with the use of hardware and software technologies for resourcefully storing huge amount of medical records. For experimentation, machine learning algorithms are applied on the data which can be quickly retrieved any time with the help of computer processing systems. It is proved from literature study that each intelligent technique has its own significance in providing inclusive information as per the scalability and diversity of data. Each classifier follows unique steps for data processing and computation which makes them distinct in producing results. Hence, to interpret multifaceted dataset, to avoid clinical inexperience and to reduce the evaluation time; this chapter proposes an intelligent medical decision support system for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. The system is built using integration of data clustering and classification performed by enhanced hierarchical clustering and random decision forest algorithms respectively. The chapter deployed individual and integrated classification methods which include random decision forest, improved random decision forest, hierarchical clustering with random decision forest, hierarchical clustering with improved random decision forest, and enhanced hierarchical clustering with improved random decision forest. Performance of all aforesaid models is compared in terms of accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. Simulation results showed that enhanced hierarchical clustering with improved random decision forest based intelligent integrated approach achieved better prediction outcomes than other individual and integrated models. In addition to higher accuracy rates, the model also attained remarkable precision and true positive rates.

7.2 Methodology

The study aims to propose an intelligent medical decision support system based on hierarchical clustering and random forest algorithms for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. The diagnostic models developed in this

work are represented as RF, ERF, HC-RF, HC-ERF and EHC-ERF. RF stands for random decision forest with classification and regression tree algorithm as the learning model. ERF indicates improved random forest algorithm with random decision tree as the learning model. HC signifies hierarchical clustering algorithm with euclidean distance function and EHC denotes enhanced hierarchical clustering which used improved distance function. HC-RF and HC-ERF indicate integration of hierarchical clustering with RF and ERF respectively. EHC-ERF symbolizes the integration of enhanced hierarchical clustering with ERF which is the best diagnostic model among all aforesaid methods. Figure 7.1 illustrates the block diagram of proposed intelligent integrated model for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. Firstly, the hepatobiliary disorder data is taken as input in form of raw instances. The data incorporates five hundred and thirty six instances with nine attributes and four target classes which are randomized using a predefined class. Then the sample values are converted from numeric to nominal format for giving input to the system. Secondly, enhanced hierarchical clustering algorithm is deployed to cluster the data. Then, the improved random forest algorithm with random decision tree as the learning model is used to predict alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. Advantages of deploying hierarchical clustering and random decision forest algorithms include small cluster generation for better prediction, efficient handling of input variables, internal unbiased estimate of generalization error, deduction of key variables in classification, resistance to over training and no apriori information needed about number of clusters. Owing to enhanced hierarchical clustering, the random decision forest method predicts hepatobiliary disorder cases efficiently. Description of intelligent clustering and classification algorithms used in the proposed classification model EHC-ERF are as follows.

Hierarchical clustering algorithm represents information by tree of clusters or by grouping data objects into hierarchy. Its structure is more instructive than the unstructured set of clusters returned by flat clustering. There is no apriori information needed about number of clusters required. It develops a sequence of nested clusters and the range is from individual clusters of single points to all-together cluster [145], [180], [181]. This sequence of nested clusters is graphically represented by dendrogram. Dendrogram is a process by which objects are grouped together or partitioned step-by-step. Let’s assume, a set of M data points (data objects) and M*M distance or similarity matrix is given. Each item is assigned to a cluster. For M number of items, M clusters are formed. It finds the nearest cluster and joins them into

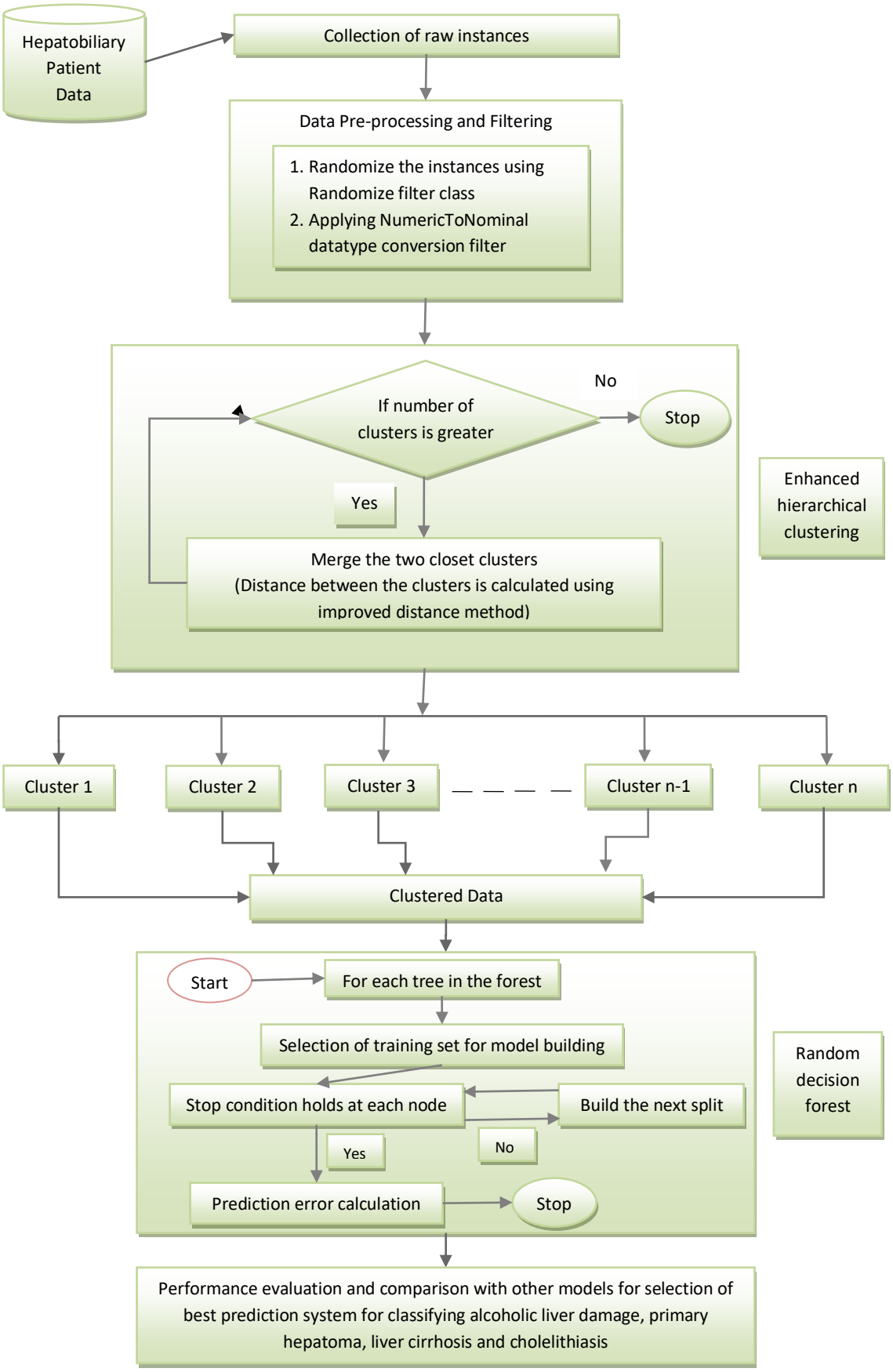

Figure 7.1. Block diagram of proposed intelligent integrated model for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis

a new single cluster. This decreases one cluster each time. Then it calculates distances or similarities between new cluster and each of old clusters. This process is repeated until there is only single cluster of size M*M is left. Before performing any clustering, it determines the proximity matrix which contains distance between each point using distance function. The procedure is represented below:

A set Y of objects $\{y_{1,\ldots\ldots}y_m\}$ $//set\ of\ data\ points$

A distance function $dist\ (k_1, k_2)$

for j=1 to m

$k_j = \{y_j\}$

end for

K = $\{k_{1,\ldots\ldots}k_m\}$ $//set\ of\ clusters$

$p = m + 1$ $//\ sequence\ number$

while K.size > 1 do

- (k_{min1}, k_{min2}) = minimum $dist(k_j, k_i)$ for all k_j, k_i in K
- remove k_{min1} and k_{min2} from K
- add $\{k_{min1}, k_{min2}\}$ to K
- $p = p+1$

end while

where euclidean function given in equation (7.1) is used to compute the distance in hierarchal clustering and an improved distance function given in equation (7.2) is used in enhanced hierarchical clustering. For instance, the euclidean distance (d) between vectors $p = p_1, p_2, \ldots p_n$ and $q = q_1, q_2, \ldots q_n$ in n space is represented as:

$$d = \sqrt{\sum_{j=1}^{n} (p_j - q_j)^2} \tag{7.1}$$

and, the improved distance between vectors $p = p_1, p_2, \ldots p_n$ and $q = q_1, q_2, \ldots q_n$ in n space is as follows:

$$d = \sqrt{(v)^{-1} \sum_{j=1}^{n} (p_j - q_j)^2} \tag{7.2}$$

where v denotes weight, $\bar{p}$ indicates mean of attributes and v is computed using the formula: $v = \dfrac{\sum_{j=1}^{n}(p_j - \bar{p})^2}{n-1}$

Random forest algorithm constructs number of decision trees at training time and returns the output of class is based on prediction of individual trees. It is an ensemble based learning

which is capable of performing both regression and classification tasks. The basic principle behind this classifier is forming a strong learner by a group of weak learners. It has the capability to create accurate classifiers by generating right kind of randomness. It resolves the problem of high bias and variance by finding average between two extremes [182]–[184]. Random forest formed with random input selection is called Forest-RI. Occurrence of forest error rate is dependent on two factors: first is correlation and second is strength of each individual tree. Correlation is directly proportional to forest error rate and strength is inversely proportional to forest rate. A tree acts as a strong classifier where error rate is low. Each tree is grown as per the following steps. In step 1, take *M* and *N* which represent number of training cases and number of variables in classifier respectively. Step 2 finds a decision at node of tree, *n* of input variables are used where *n*<*N*. In step 3, training set for tree is picked *m* times with substitution from *M* training cases that are accessible. By predicting their classes, left cases are utilized to estimate the error of tree. In step 4, *n* factors are arbitrarily picked for every node of tree on which to make the choice at that node. On the basis of *n* variables presented in training data, calculate the finest split. Finally in step 5, each tree is grown to the maximum extent and there is no pruning. For predicting a new instance, the tree is traversed from top to bottom and then assigned a label associated with the training terminal node. This process is iterated over all trees and the random forest classifier is obtained with majority vote among these classification trees. For instance, the hepatobiliary training data is represented as $D_m = (Y_1, Z_1), \dots\dots, (Y_m, Z_m)$ where Y and Z are independent random variables which are same as the autonomous sample pair (Y, Z). This training set D_m is used to give estimation of $f_m : [0,1]^k \to R$ of function f. Mean square error f_m is consistent if $H[f_m(Y) - f(Y)]^2 \to 0$ as $m \to \infty$. Input random vector $y \in [0,1]^k$, the aim is to predict response $Z \in R$ by regression function approximation, i.e. $f(y) = H[\mathrm{Z}|\mathrm{Y} = \mathrm{y}]$. Random forest predictor consists of F randomized regression trees. The value predicted at query point y for p^{th} tree in family is actually denoted by $f_m(y; \theta_p, D_m)$ where $\theta_1, \dots\dots, \theta_f$ are independent random variables. Before growing of individual trees, θ is used to resample the training data and to select the consecutive directions for partitioning. At this stage, different trees are combined to make finite forest estimate.

$$f_{F,m}(y; \theta_1, \dots\dots, \theta_f, D_m) = \frac{1}{F} \sum_{p=1}^{F} f_m(y; \theta_p, D_m) \tag{7.3}$$

Since *F* may be chosen randomly high then let's assume *F* tends to infinity and the forest estimate is denoted as:

$$f_{\infty,m}(y; D_m) = H_\theta \left[f_m(y; \theta_p, D_m) \right] \quad (7.4)$$

Here, H_θ denotes probability with respect to arbitrary factor θ which is conditional on D_m. The process "$F \to \infty$" is acceptable by large numbers and is conditional on D_m.

$$\lim_{F \to \infty} f_{F,m}\left(y; \theta_1, \dots\dots, \theta_f, D_m\right) = f_{\infty,m}(y; D_m) \quad (7.5)$$

In classification, response variable Z takes value in range [0,1] and the value of Z is calculated with known variable Y. Classifier f_m is a measureable function of $y \; and \; D_m$ and the label of Z is also approximated from $y \; and \; D_m$. Classification and regression tree is used as the learning model in random forest algorithm and random decision tree is used as the learning model in improved random decision forest. The classifier f_m is said to be consistent if conditional possibility of error $E(f_m) = K[f_m(Y) \neq Z|D_m]$ satisfies $\lim_{m \to \infty} HE(f_m) = E^*$ where E^* is an unknown error but optimal bayes classifier is:

$$f^*(y) = \begin{cases} 1 & if \; K[Z = 1|Y = y] > K[Z = 0|Y = y] \\ 0 & otherwise \end{cases} \quad (7.6)$$

The random forest classifier is obtained with majority vote among classification trees, i.e.

$$f_{F,m}\left(y; \theta_1, \dots\dots, \theta_f, D_m\right) = \begin{cases} 1 \; if \; \frac{1}{F}\sum_{p=1}^{F} f_m(y; \theta_p, D_m) > 1/2 \\ 0 \qquad\qquad otherwise \end{cases} \quad (7.7)$$

7.3 Results and Discussion

The hepatobiliary disorder dataset obtained from a university-affiliated hospital in Japan is used for experimentation. The dataset includes nine attributes (continuous real-valued measurements from biomedical test), four classes, and five hundred and thirty six instances. Attributes contain information about glutamic oxalacetic transaminase, glutamic pyruvic transaminase, lactate dehydrogenase, gamma glutamyl transpeptidase, blood urea nitrogen, mean corpuscular volume of red blood cells, mean corpuscular hemoglobin, total bilirubin and creatinine. Four target classes include alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. Each instance in the data represents information of a single male or female. The dataset is randomly split into training set containing seventy percent of data and testing set containing remaining thirty percent. This division validates the proposed diagnostic model and reduces the biasness associated with instances. Obtained results of the developed individual and integrated classification models are compared using accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. Principally the output results of a classification model are produced in the form of TP,

TN, FP and FN; and then the aforesaid parameters are calculated using these values. TP indicates true positive (diseased people correctly recognized as diseased), TN is true negative (normal people correctly recognized as normal), FN is false negative (diseased people incorrectly identified as normal), and FP expresses false positive (normal people incorrectly identified as diseased). Accuracy is the ability to distinguish target classes correctly. It is calculated using the ratio of sum of all TP and TN to sum of all TP, TN, FP and FN. True positive rate is also known as sensitivity or recall which measures the proportion of instances that are correctly classified as class A, among all truly class A instances. It is computed using the ratio of TP to sum of TP and FN. Precision is also known as positive predictive value which measures the proportion of instances that truly belong to class A, among all classified class A instances. It is calculated using ratio of TP to sum of TP and FP. F-measure is also known as F-score which computes performance of a model for positive class. It is calculated using the ratio of multiplication of both precision and recall with 2 to sum of precision and recall. Kappa statistic computes the agreement of prediction with true class. Agreement is scaled between 0.0 and 1.0 where the later value signifies complete agreement. Mean absolute error is an average of absolute errors which is not squared before averaging and it is used to quantify the closeness of predictions to the eventual outcomes. Unlike MAE, root mean squared error squares the difference between predictions and eventual outcomes before averaging absolute errors in order to assign more weight to large errors. Table 7.1 details the description of biomedical test attributes and their measurement unit.

Table 7.1: The attributes of hepatobiliary disorder dataset

Attribute	Description	Unit measurement
GOT	Glutamic oxaloacetic transaminase	Karmen unit
GPT	Glutamic pyruvic transaminase	Karmen unit
LDH	Lactate dehydrase	iu/l
GGT	Gamma glutamyl transpeptidase	µ/ml
BUN	Blood urea nitrogen	mg/dl
MCV	Mean corpuscular volume of red blood cells	fl
MCH	Mean corpuscular hemoglobin	pg
TBIL	Total bilirubin	mg/dl
CRTNN	Creatinine	mg/dl

The intelligent diagnostic approaches build for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis are represented as RF, ERF, HC-RF, HC-ERF, and EHC-ERF. RF stands for random forest algorithm, ERF signifies improved random forest algorithm, HC-RF indicates integration of hierarchical clustering with RF, HC-ERF

stands for integration of hierarchical clustering with ERF, and EHC-ERF symbolizes the integration of enhanced hierarchical clustering with ERF. Figure 7.2, 7.3, 7.4, 7.5, 7.6, 7.7 and 7.8 illustrate the performance comparison among build classification models using accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error rates respectively.

Figure 7.2 depicts that RF had 85.71% accuracy, ERF had 86.96% accuracy, HC-RF had 91.3% accuracy, HC-ERF had 93.79% accuracy and EHC-ERF had 96.27% accuracy. Figure 7.3 shows that RF had 85.7% true positive rate, ERF had 87% true positive rate, HC-RF had 91.3% true positive rate, HC-ERF had 93.8% true positive rate and EHC-ERF had 96.3% true positive rate. Figure 7.4 portrays that RF had 86.9% precision, ERF had 87.7% precision, HC-RF had 91.1% precision, HC-ERF had 93.8% precision and EHC-ERF had 96.4% precision. Figure 7.5 describes that RF had 86% f-measure, ERF had 87% f-measure, HC-RF had 91.1% f-measure, HC-ERF had 93.6% f-measure and EHC-ERF had 96.1% f-measure. Figure 7.6 represents that RF had 80.92% kappa statistic, ERF had 82.57% kappa statistic, HC-RF had 76.27% kappa statistic, HC-ERF had 82.75% kappa statistic and EHC-ERF had 88.23% kappa statistic. Figure 7.7 depicts that RF had 13.52% mean absolute error, ERF had 12.41% mean absolute error, HC-RF had 6.04% mean absolute error, HC-ERF had 6.27% mean absolute error and EHC-ERF had 5.99% mean absolute error. Figure 7.8 presents that RF had 24.68% root mean squared error, ERF had 22.17% root mean squared error, HC-RF had 19.56% root mean squared error, HC-ERF had 14.74% root mean squared error and EHC-ERF had 14.9% root mean squared error.

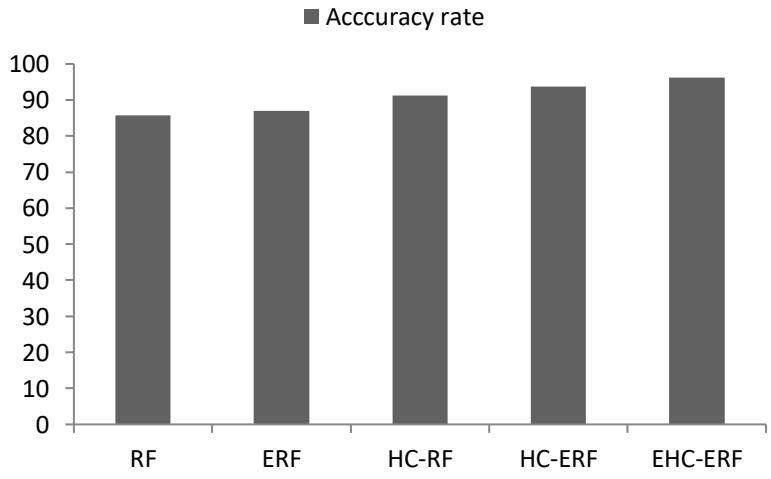

Figure 7.2: The comparative view of obtained accuracy rates

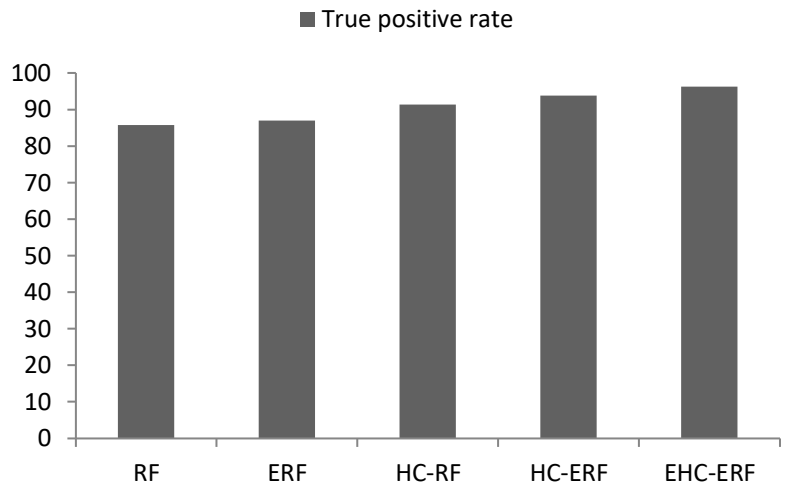

Figure 7.3: The comparative view of obtained true positive rates

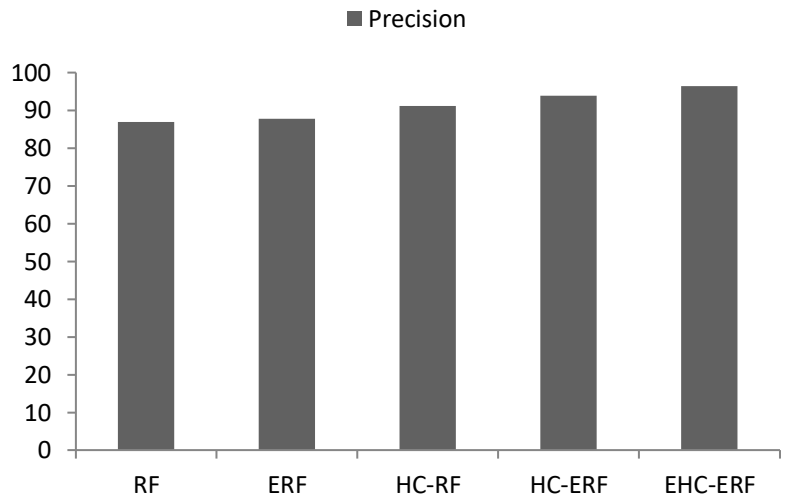

Figure 7.4: The comparative view of obtained precision rates

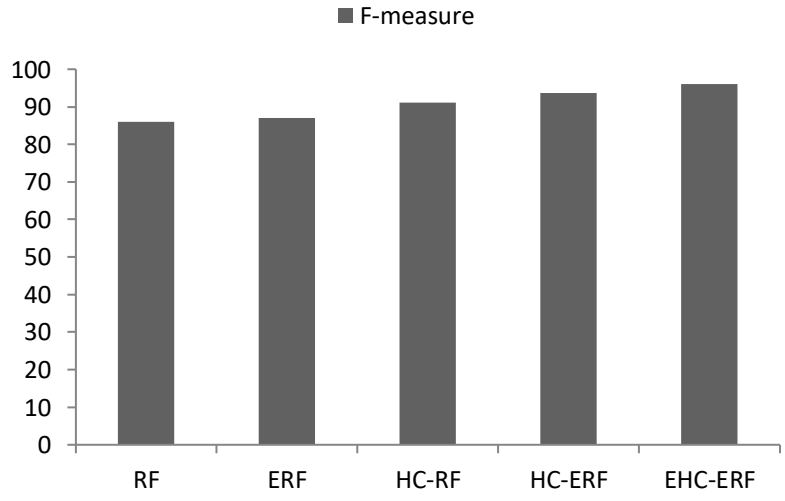

Figure 7.5: The comparative view of obtained F-measure rates

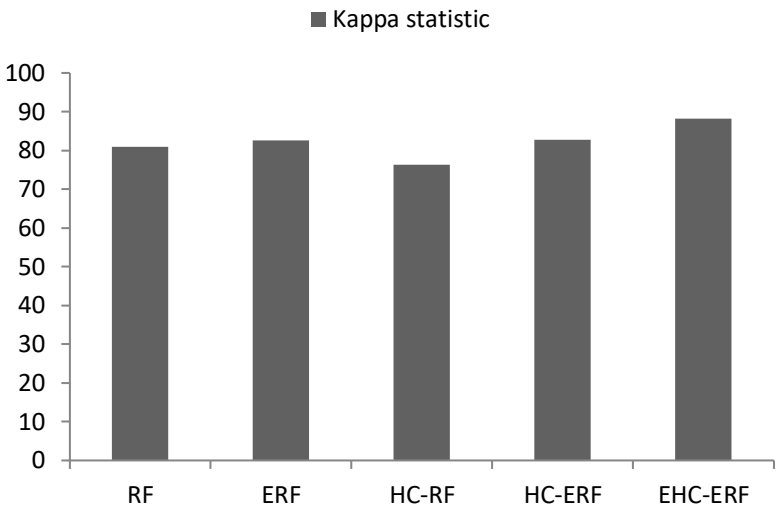

Figure 7.6: The comparative view of obtained kappa statistic

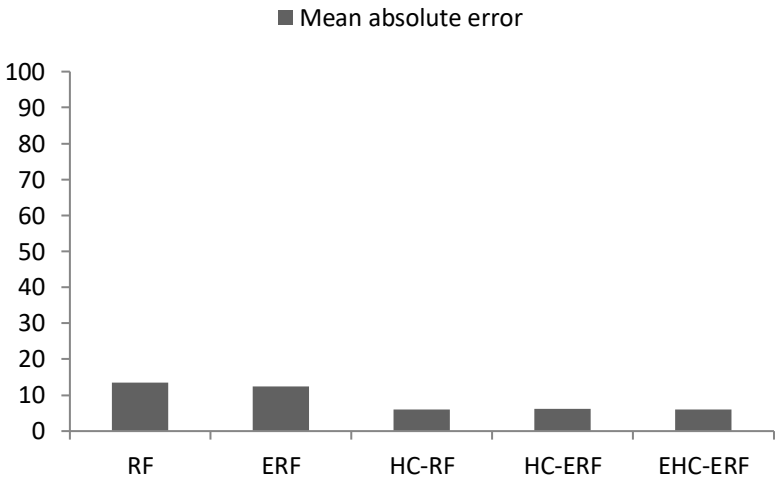

Figure 7.7: The comparative view of obtained mean absolute error rates

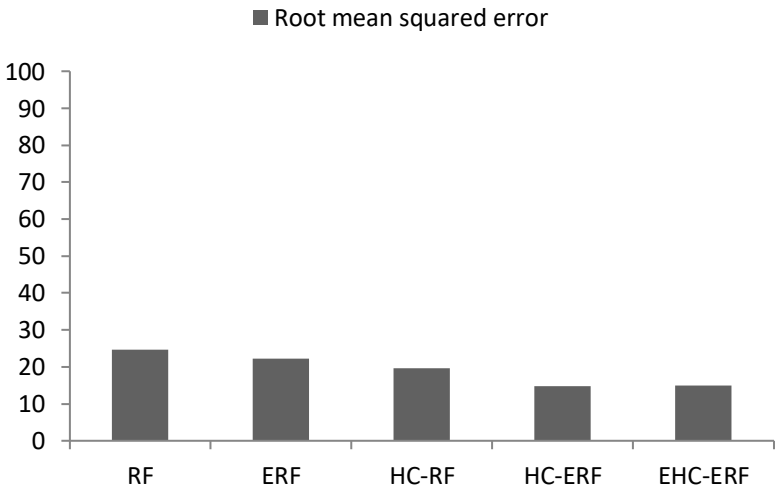

Figure 7.8: The comparative view of obtained root mean squared error rates

To select the most efficient medical decision support system for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis; obtained results (accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error rates) of all proposed models are compared (Table 7.2). Prediction results of EHC-ERF are also compared to other hepatobiliary classification methods mentioned in the literature. Y. Hayashia et al. [36] stated that LDA, Fuzzy neural network, NeuroRule and NeuroLinear achieved accuracy rates of 63.2%, 77.3%, 88.3% and 90.2% respectively. In FNN, the back propagation neural network model is applied where the input data is in the form of fuzzy arithmetic and fuzzy numbers. S.K. Pal and S. Mitra [185] mentioned that fuzzy multilayer perceptron network attained 76.0% and 88.9% accuracies for best and second best choice criteria where the combination of membership values is given as input to MLP in the set categorization as low, medium and high. The fuzziness incorporated enhanced neural network weights through backpropagating the errors. Y. Hayashi and R. Setiono [39] mentioned that average accuracy rates of 30, 5, 10 and 15 neural networks are 90.27%, 90.92%, 91.78% and 91.92%; average accuracy rates of developed biased neural networks are 92.64%, 92.02%, 93.25% and 94.48%; average accuracy rates of applying neural networks as second level model are 87.73%, 90.18%, 84.66%, 87.12%, 91.41%, 88.34% and 89.57%. L. K. Ming et al. [60] presented a fuzzy model based on enhanced supervised fuzzy clustering algorithm where global k-means method is used to initialize the fuzzy model. This method overcomes the limitation of simple k-means i.e. unknown number of clusters and random generation of initial positions of clusters. Supervised fuzzy clustering with random initialization had 58.57% accuracy and enhanced supervised fuzzy clustering with global k-means had 58.78% accuracy. A. Niyom, et al. [186] obtained 76.99% accuracy using neural network classification; K. Oaba and E. Tazaki [187] achieved 77% extended genetic algorithm with neutral mutation-application; Y. Hayashi, et al. [188] attained 88.3% using rules extracted from artificial neural networks; T. Ichimura, et al. [189] stated 93% using fuzzy rules and neural networks; and T. Ichimura, et al. [190] mentioned 87.9% using an adaptive evolutional neuro-learning method with genetic search. Experimental results showed that RF and ERF based models have not shown significant prediction performance. Although, HC-RF and HC-ERF attained enhanced accuracy rates than the aforesaid models but EHC-ERF achieved highest among all and is selected as the best classification model. The proposed system also outperforms methods developed in the literature. The intelligent integrated approach combines advantages of hierarchical clustering and random decision forest such as enhanced prediction results through generation of smaller clusters, consistency of cluster results on

different algorithms runs, precise learning, estimation of key variables, fine computation of proximities between pairs of cases and no apriori information required about cluster numbers. Undoubtedly clinicians play a prime role in final judgment on patient health condition but carrying out a resourceful diagnosis is an intricate job that requires enormous medical experience. Certainly, these computationally intelligent medical support systems cannot replace physicians' role but may positively assist them in examining medical records by acting as a second opinion. This chapter is also an effort in that direction which proposed an enhanced hierarchical clustering and improved random decision forest algorithm based predictive model for the efficient classification of alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis.

Table 7.2: The simulation results of intelligent integrated models

Classification model	RF	ERF	HC-RF	HC-ERF	EHC-ERF
Accuracy	85.71%	86.96%	91.3%	93.79%	96.27%
TPR	85.7%	87%	91.3%	93.8%	96.3%
Precision	86.9%	87.7%	91.1%	93.8%	96.4%
F-measure	86%	87%	91.1%	93.6%	96.1%
Kappa statistic	80.92%	82.57%	76.27%	82.75%	88.23%
MAE	13.52%	12.41%	6.04%	6.27%	5.99%
RMSE	24.68%	22.17%	19.56%	14.74%	14.9%

7.4 Conclusions

Diagnosing a disease is one of the most difficult responsibilities clinicians' do have as one minute error can endanger patient life. Implementation of intelligent techniques has done a major transformation in predicting health examination data, and the medical domain has also been widely affected by this renovation. Classification of alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis disease is also an intricate task. As a part of constant efforts for making hepatobiliary disorder classification process well-organized and proficient, this chapter built an intelligent integrated model based on enhanced hierarchical clustering and improved random decision forest algorithms. The model has advantages of both hierarchical clustering and random decision forest such as enhanced prediction results through generation of smaller clusters, consistency of cluster results on different algorithms runs, precise learning, estimation of key variables, fine computation of proximities between pairs of cases, and no apriori information required about cluster numbers. The integrated approach showed capability of improving complex medical decisions through clustered data.

The prediction was carried out using a data of five hundred and thirty six cases of hepatobiliary disorder. Simulation results confirmed the superiority of proposed approach to other diagnostic models implemented in the chapter and mentioned in literature as well. Mean absolute error and root mean squared error rates were also small. Thousands people lose their lives because of erroneous evaluation and inappropriate treatment of alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis as the medical cases are still largely influenced by subjectivity of physicians. The proposed medical decision support system can be applied as a liver specialist assistant or as a model to train novice medical students. The system will also help physicians in evaluating complex cases that are otherwise hard to perceive. It also shows the capability of non-invasive method and to reduce the need of liver biopsy to a possible extent.

The findings of the chapter have been published in the Journal of Healthcare Engineering, Volume 2018, Article ID 1469043, 9 pages, 2018. DOI:10.1155/2018/1469043, mentioned under the list of publications at the end of the chapter 8.

Chapter 8

Conclusion and Future Scope

This chapter is organized as follows: Section 8.1 summarizes the work described in the thesis by empahsising on the major contributions of proposed intelligent models in liver disease diagnosis and Section 8.2 discusses some areas of future research.

8.1 Summary of Deductions

The thesis was set out to study and develop intelligent techniques based computational models for the diagnosis of liver disease. Implementation of these models has contributed a major transformation in the field of information retrieval, and the medical domain has also been widely affected by this renovation. Intelligent techniques imitate diagnostic systems to work like human brain. In this research work, the intelligent models were proposed for identifying liver disease, predicting degree of liver damage, classifying primary biliary cirrhosis, diagnosing hepatitis disease, and classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. Chapter 2 presented a comprehensive literature review on intelligent techniques applied to liver disease. Different types of liver disease covered in the chapter were hepatitis, liver fibrosis, liver cancer, fatty liver, general liver damage, alcoholic liver damage, primary hepatoma, cholelithiasis and liver cirrhosis. Chapter 3 introduced an efficient diagnosis system for detection of liver disease using a novel integrated method based on principal component analysis and correlation distance metric based k-nearest neighbor where the former was used for feature extraction and the later was deployed for classification. Experimental results showed that PCA-LDA, PCA-DLDA, PCA-QDA and PCA-DQDA based methods had not shown significant diagnostic performance. Although, PCA-LSSVM showed enhanced accuracy rates then aforesaid models but PCA-KNN achieved highest among all. The comparison was done using accuracy, sensitivity, specificity, PPV, and NPV rates. Transformation of features into new space by PCA made the classifier more efficient. This intelligent integrated approach combines advantages of both PCA and KNN such as high classification rates, good generalization, plain structure and efficient problem solving ability through feature extraction.

Chapter 4 explained the implementation of dimensionality reduction and classification algorithms for the prediction of primary biliary cirrhosis stages. Dimensionality reduction techniques were divided into feature extraction and feature selection. The chapter introduced two intelligent integrated models. In feature selection based computational model; first kullback–leibler divergence technique was implemented for feature ranking and then least squares support vector machine approach was deployed to classify primary biliary cirrhosis stages. In feature extraction based computational method; first principal component analysis was used for extraction, and then euclidean distance and nearest rule based k-nearest neighbor algorithm was employed for predicting primary biliary cirrhosis stages. It was observed from experiments that KLD-LSSVM in feature selection category and PCA-KNN in feature extraction category outperformed all other classifiers and were selected as best predictive models for PBC stages. The comparison was done using accuracy, sensitivity, specificity, PPV, and NPV rates. KLD based integrated method incorporated features given in table 4.8 for classifying stage 1, 2, 3 and 4. Variation in ranking of features with respect to PBC stages indicated the complex nature of diagnosis process. For instance; ascites, hepato, spiders, edema, BIL, protime, trig, chol, copper, AST, ALB and ALK were the key features for stage 1 prediction. Ascites, edema, protime, copper, spiders, hepato, BIL, ALB, platelet, chol, trig and AST were vital of stage 2. Ascites, chol, edema, protime, ALB, ALK, copper, trig, age, platelet, hepato and gender were crucial for stage 3. Ascites, edema, hepato, copper, ALB, protime, spiders, chol, BIL, platelet, ALK and trig were essential for stage 4. Results also showed that exclusion of age, gender and platelet count features did not lessen the prediction accuracy of stage 1. Likewise, leaving out age, gender and alkaline phosphotase in stage 2; removal of spiders, serum bilirubin and aspartate aminotransferase in stage 3; and elimination of age, gender and aspartate aminotransferase in stage 4 showed no decline in identifying PBC stages. Presence of ascites and edema were the most common and influential attributes in all four stages. As the disease progresses, amount of cooper in urine starts increasing which indicates a sign of being in the end stage of PBC. Primary biliary cirrhosis is a general cause of liver cirrhosis globally and will surely keep engaging the novice researchers and physicians in its assessment. Chapter 5 demonstrated the effectiveness of euclidean distance function based k-nearest neighbor computational model for assessing degree of liver damage. Learning process in KNN is almost zero whose simplest variation is 1NN. Here the value of K is equal to 1. It aims to find the nearest neighbor *n* based on the training data. KNN solves the problem of over-fitting by searching multiple samples as nearest neighbor. Simulations results showed that euclidean distance based k-nearest neighbor model achieved better prediction outcomes

than other classifiers implemented in the study. The comparison was done using accuracy, sensitivity, specificity, PPV, and NPV rates. In addition to higher accuracy rates, it also attained remarkable sensitivity and specificity which is a challenging task given an uneven variance among attribute values. The method has given best decisions in shortest time possible and was considered more accurate as the technique used is data driven. The model filters overflow of information, data and knowledge. It provides ability to learn from N number of experiences with respect to some task T, and to act appropriately in an uncertain environment for increasing the probability of success. KNN makes the proposed model self modifying and highly automated which continues to improve over time as it learns with more data.

Chapter 6 introduced an intelligent hybrid approach for hepatitis disease diagnosis by combining enhanced k-means clustering and ensemble learning algorithms. The advantage of deploying ensemble learning was that it constructed a set of hypothesis using multiple learners for solving hepatitis cases. The proposed model worked with combination of data clustering and ensemble learning performed by enhanced k-means clustering; and improved adaptive boosting, bagged decision tree and J48 decision tree respectively. The diagnostic approaches implemented in the study were represented as EL1, EL2, KM-EL1, KM-EL2, KM-EL3, KM-EL4 and KM1-EL4. EL1 stands for ensemble learning with REP tree as base learner and ZeroR as meta-classifier. EL2 signifies ensemble learning with J48 decision tree as base learner and random forest as meta-classifier. EL3 indicates ensemble learning with adaptive boosting as base learner and J48 decision tree as meta-classifier where learning model of adaptive boosting was decision stump algorithm. EL4 stands for enhanced ensemble learning with improved adaptive boosting and bagged decision tree as base learners and J48 decision tree as meta-classifier where learning model of adaptive boosting was random forest algorithm. KM stands for k-means clustering and KM1 denotes enhanced k-means clustering. KM-EL1, KM-EL2, KM-EL3 and KM-EL4 indicate integration of k-means clustering with EL1, EL2, EL3 and EL4 respectively. KM1-EL4 symbolizes the integration of enhanced k-means clustering and EL4 which was the best diagnostic approach among all aforesaid methods. Performance was compared in terms of accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. The integrated model combines advantages of k-means clustering, adaptive boosting, bagged decision tree and J48 algorithms such as high classification rates, good generalization, plain structure and efficient problem solving ability. Chapter 7 presented a new intelligent medical decision

support system based on enhanced hierarchal clustering and random decision forest for classifying alcoholic liver damage, primary hepatoma, liver cirrhosis and cholelithiasis. The system was built using integration of data clustering and classification performed by enhanced hierarchical clustering and random decision forest algorithms respectively. The diagnostic models developed in the study were represented as RF, ERF, HC-RF, HC-ERF and EHC-ERF. RF stands for random decision forest with classification and regression tree algorithm as the learning model. ERF indicates improved random forest algorithm with random decision tree as the learning model. HC signifies hierarchical clustering algorithm with euclidean distance function and EHC denotes enhanced hierarchical clustering which used improved distance function. HC-RF and HC-ERF indicate integration of hierarchical clustering with RF and ERF respectively. EHC-ERF symbolizes the integration of enhanced hierarchical clustering with ERF which was the best diagnostic model among all aforesaid methods. Performance was compared in terms of accuracy, true positive rate, precision, f-measure, kappa statistic, mean absolute error and root mean squared error. The model combines advantages of hierarchical clustering and random decision forest such as enhanced prediction results through generation of smaller clusters, consistency of cluster results on different algorithms runs, precise learning, estimation of key variables, fine computation of proximities between pairs of cases and no apriori information required about cluster numbers.

In conclusion, it can be said that each intelligent technique has the capability to outperform other techniques based on the type and structure of data. Raw data faces various challenges that make traditional methods improper for knowledge extraction. Deployment of ITs is a powerful method to extract knowledge from medical data and this kind of extracted knowledge is used as a new knowledge. ITs can be applied to both linear and non-linear data. In medical field most of the data collected is non-linear in nature and hence these techniques are efficient for medical applications. ITs are immune to noise present in data. Medical data is prone to have noise and hence again their deployment is beneficial. In medical field, it is not possible to collect data on regular bases. If data is collected from patients and some readings may be missing then ITs can deal with those missing data without creating any error in output. ITs applied correctly, can give near to accurate answer even when data is not sufficient. ITs also have the capability to get rid of unwanted attributes for better results.

8.2 Future Scope of Work

Artificial intelligence in medicine is an exceptionally dynamic research domain. Intelligent techniques have been continuously improved or combined with other methods for the efficient prediction of disease. Likewise, the proposed systems in this work can also be extended for predicting other types of liver disease like non alcoholic fatty liver disease, neonatal hepatitis, primary sclerosing cholangitis and wilson disease. Dimensionality reduction methods can be further explored to remove multi-collinearity in medical records which eventually improves the performance of computational models. It may be appropriate to use large databases to enhance the system learning capability and robustness. New intelligent approaches can be developed for solving general challenges in medical data such as skewness, high dimensionality, missing value, large volume, noise and heterogeneity. Another potentially productive area of research might be the use of developed computational models for hardware implementation.

List of Publications

1. Aman Singh and Babita Pandey, A New Intelligent Medical Decision Support System Based on Enhanced Hierarchical Clustering and Random Decision Forest for the Classification of Alcoholic Liver Damage, Primary Hepatoma, Liver Cirrhosis, and Cholelithiasis, Journal of Healthcare Engineering, Volume 2018, Article ID 1469043, 9 pages, 2018. DOI:10.1155/2018/1469043

2. Aman Singh and Babita Pandey, A KLD-LSSVM based computational method applied for feature ranking and classification of primary biliary cirrhosis stages, International Journal of Computational Biology and Drug Design, Vol. 10, No. 1, pp. 24-38, 2017, Inderscience Publishers, United Kingdom, DOI: 10.1504/IJCBDD.2017.082807.

3. Aman Singh and Babita Pandey, An Efficient Diagnosis System for Detection of Liver Disease Using a Novel Integrated Method Based on Principal Component Analysis and K-Nearest Neighbor, International Journal of Healthcare Information Systems and Informatics, Vol. 11, No. 4, pp. 56-69, 2016, IGI Global Publishing, United States, DOI: 10.4018/IJHISI.2016100103.

4. Aman Singh and Babita Pandey, Diagnosis of Liver Disease Using Correlation Distance Metric Based K-Nearest Neighbor Approach, International Symposium on Intelligent Systems Technologies and Applications (ISTA), September 21–24, 2016, Jaipur, India, Advances in Intelligent Systems and Computing book series (AISC), Vol. 530, pp. 845-856, 2016, Springer International Publishing, DOI: 10.1007/978-3-319-47952-1.

5. Aman Singh and Babita Pandey, Diagnosis of Liver Disease by Using Least Squares Support Vector Machine Approach, International Journal of Healthcare Information Systems and Informatics, Vol. 11, No. 2, pp. 62-75, 2016, IGI Global Publishing, United States, DOI: 10.4018/IJHISI.2016040104.

6. Aman Singh and Babita Pandey, An euclidean distance based KNN computational method for assessing degree of liver damage, International Conference on Inventive Computation Technologies (ICICT), August 26-27, 2016, Coimbatore, India, Vol. 1, IEEE, DOI: 10.1109/INVENTIVE.2016.7823222.

7. Aman Singh and Babita Pandey, Classification of primary biliary cirrhosis using hybridization of dimensionality reduction and machine learning methods, International Conference on Inventive Computation Technologies (ICICT), August 26-27, 2016, Coimbatore, India, Vol. 1, IEEE, DOI: 10.1109/INVENTIVE.2016.7823222.

8. Aman Singh and Babita Pandey, Liver disorder diagnosis using linear, nonlinear and decision tree classification algorithms, International Journal of Engineering and Technology, Vol. 11, No. 4, pp. 2059-2069, 2016, Engg Journals Publications, India, DOI: 10.21817/ijet/2016/v8i5/160805424.

9. Sakshi Takkar, Aman Singh and Babita Pandey, Application of Machine Learning Algorithms to a Well Defined Clinical Problem-Liver Disease, International Journal of E-Health and Medical Communications, Vol. 8, No. 4, 2017, IGI Global Publishing, United States.

10. Aman Singh and Babita Pandey, Intelligent techniques and applications in liver disorders: a survey, International Journal of Biomedical Engineering and Technology, Vol. 16, No. 1, pp. 27-70, 2014, Inderscience Publishers, United Kingdom, DOI: 10.1504/IJBET.2014. 065638.

References

[1] B. N. Li, C. K. Chui, S. Chang, and S. H. Ong, "A new unified level set method for semi-automatic liver tumor segmentation on contrast-enhanced CT images," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9661–9668, 2012.

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

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

[4] İ. Ö. Bucak and S. Baki, "Diagnosis of liver disease by using CMAC neural network approach," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6157–6164, 2010.

[5] A. F. Peery, S. D. Crockett, A. S. Barritt, E. S. Dellon, S. Eluri, L. M. Gangarosa, E. T. Jensen, J. L. Lund, S. Pasricha, T. Runge, M. Schmidt, N. J. Shaheen, and R. S. Sandler, "Burden of Gastrointestinal, Liver, and Pancreatic Diseases in the United States," *Gastroenterology*, vol. 149, no. 7, p. 1731–1741e3, 2015.

[6] Canadian Liver Foundation, "Liver Disease in Canada: A crisis in the making," *Can. Liver Found.*, no. March, pp. 1–70, 2013.

[7] M. Blachier, H. Leleu, M. Peck-Radosavljevic, D.-C. Valla, and F. Roudot-Thoraval, "The burden of liver disease in Europe: a review of available epidemiological data.," *J. Hepatol.*, vol. 58, no. 3, pp. 593–608, 2013.

[8] S. Kalra, M. Vithalani, G. Gulati, C. M. Kulkarni, Y. Kadam, J. Pallivathukkal, B. Das, R. Sahay, and K. D. Modi, "Study of prevalence of nonalcoholic fatty liver disease (NAFLD) in type 2 diabetes patients in India (SPRINT).," *J. Assoc. Physicians India*, vol. 61, no. 7, pp. 448–453, 2013.

[9] W. Dunn, R. Xu, D. D. L. Wingard, C. Rogers, P. Angulo, Z. M. Z. Younossi, J. B. Schwimmer, and S. Jb, "Suspected nonalcoholic fatty liver disease and mortality risk in a population-based cohort study.," *Am. J. Gastroenterol.*, vol. 103, no. 9, pp. 2263–2271, 2008.

[10] P. Paschos and K. Paletas, "Non alcoholic fatty liver disease and metabolic syndrome," *Hippokratia*, vol. 13, no. 1, pp. 9–19, 2009.

[11] M. Obika and H. Noguchi, "Diagnosis and evaluation of nonalcoholic fatty liver disease.," *Exp. Diabetes Res.*, vol. 2012, p. 145754, Jan. 2012.

[12] "Global Burden Of Liver Disease: A True Burden on Health Sciences and Economies,

World Gastroenterology Organisation." [Online]. Available: http://www.worldgastroenterology.org/publications/e-wgn/e-wgn-expert-point-of-view-articles-collection/global-burden-of-liver-disease-a-true-burden-on-health-sciences-and-economies. [Accessed: 07-Feb-2017].

[13] "Liver Disease - The Big Picture." [Online]. Available: http://www.liverfoundation.org/education/liverlowdown/ll1013/bigpicture/. [Accessed: 07-Feb-2017].

[14] L. Ali, A. Hussain, J. Li, A. Shah, U. Sudhakr, M. Mahmud, U. Zakir, X. Yan, B. Luo, and M. Rajak, "Intelligent image processing techniques for cancer progression detection, recognition and prediction in the human liver," in *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, 2014, pp. 25–31.

[15] A. E. Ozturk and M. Ceylan, "Fusion and ANN based classification of liver focal lesions using phases in magnetic resonance imaging," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 2015, pp. 415–419.

[16] W. H. Elmasry, H. M. Moftah, N. El-Bendary, and A. E. Hassanien, "Performance evaluation of computed tomography liver image segmentation approaches," in *2012 12th International Conference on Hybrid Intelligent Systems (HIS)*, 2012, pp. 109–114.

[17] M. Tsiplakidou, M. G. Tsipouras, P. Manousou, N. Giannakeas, and A. T. Tzallas, "Automated Hepatic Steatosis Assessment through Liver Biopsy Image Processing," in *2016 IEEE 18th Conference on Business Informatics (CBI)*, 2016, pp. 61–67.

[18] W. Nanyue, Y. Youhua, H. Dawei, L. Jia, L. Tongda, S. Zengyu, C. Yanping, Xueliyuan, and Wangjia, "Comparative study of pulse-diagnosis signals between 2 kinds of liver disease patients based on the combination of unsupervised learning and supervised learning," in *2013 IEEE International Conference on Bioinformatics and Biomedicine*, 2013, pp. 260–262.

[19] R. S. Sutton and A. G. Barto, "Reinforcement learning : an introduction," *Neural Networks IEEE Trans.*, vol. 9, no. 5, p. 1054, 2013.

[20] J. H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Comput. Stat. Data Anal.*, vol. 53, no. 11, pp. 3735–3745, 2009.

[21] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection *," *Stat. Surv.*, vol. 4, pp. 40–79, 2010.

[22] A. Singh and B. Pandey, "Intelligent techniques and applications in liver disorders: a survey," *Int. J. Biomed. Eng. Technol.*, vol. 16, no. 1, pp. 27–70, 2014.

[23] B. Pandey and R. B. Mishra, "Knowledge and intelligent computing system in

medicine," *Comput. Biol. Med.*, vol. 39, no. 3, pp. 215–230, 2009.

[24] V. Mitra and J. Metcalf, "Metabolic functions of the liver," *Anaesthesia and Intensive Care Medicine*, vol. 10, no. 7. pp. 334–335, 2009.

[25] M. Kasslin, J. Kangas, and O. Simula, "Artificial Neural Networks," *Artif. Neural Networks*, pp. 1531–1534, 1992.

[26] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *J. Appl. Biomed.*, vol. 11, pp. 47–58, 2013.

[27] L. Ozyilmaz and T. Yildirim, "Artificial neural networks for diagnosis of hepatitis disease," *Proc. Int. Jt. Conf. Neural Networks, 2003.*, vol. 1, 2003.

[28] A. M. Hashem, M. E. M. Rasmy, K. M. Wahba, and O. G. Shaker, "Prediction of the degree of liver fibrosis using different pattern recognition techniques," in *2010 5th Cairo International Biomedical Engineering Conference, CIBEC 2010*, 2010, pp. 210–214.

[29] S. Içer, S. Kara, and A. Güven, "Comparison of multilayer perceptron training algorithms for portal venous doppler signals in the cirrhosis disease," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 406–413, 2006.

[30] L. Autio, M. Juhola, and J. Laurikkala, "On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension," *Comput. Biol. Med.*, vol. 37, no. 3, pp. 388–397, 2007.

[31] S. a. Azaid, M. W. Fakhr, and A. F. a. Mohamed, "Automatic Diagnosis of Liver Diseases from Ultrasound Images," *2006 Int. Conf. Comput. Eng. Syst.*, pp. 313–319, 2006.

[32] C.-C. L. C.-C. Lee, P.-C. C. P.-C. Chung, and Y.-J. C. Y.-J. Chen, *Classification of liver diseases from CT images using BP-CMAC neural network*. 2005.

[33] D. a. Elizondo, R. Birkenhead, M. Gamez, N. Garcia, and E. Alfaro, "Linear separability and classification complexity," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7796–7807, 2012.

[34] Yan Sun, Jianming Lu, and T. Yahagi, "Ultrasonographic classification of Cirrhosis based on pyramid neural network," in *Canadian Conference on Electrical and Computer Engineering, 2005.*, pp. 1682–1685.

[35] S. Ansari, I. Shafi, and A. Ansari, "Diagnosis of liver disease induced by hepatitis virus using Artificial Neural Networks," *Multitopic Conf. (INMIC), 2011 IEEE ...*, pp. 8–12, 2011.

[36] Y. Hayashi, R. Setiono, and K. Yoshida, "A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders," *Artificial Intelligence in Medicine*, vol. 20, no. 3. pp. 205–216, 2000.

[37] M. Aldape-Perez, C. Yanez-Marquez, O. Camacho-Nieto, and A. J Arguelles-Cruz, "An associative memory approach to medical decision support systems.," *Comput. Methods Programs Biomed.*, vol. 106, no. 3, pp. 287–307, 2012.

[38] P. Revesz and T. Triplet, "Classification integration and reclassification using constraint databases," *Artif. Intell. Med.*, vol. 49, no. 2, pp. 79–91, 2010.

[39] Y. Hayashi and R. Setiono, "Combining neural network predictions for medical diagnosis," *Comput. Biol. Med.*, vol. 32, no. 4, pp. 237–246, 2002.

[40] M. Rouhani and M. M. Haghighi, "The Diagnosis of Hepatitis Diseases by Support Vector Machines and Artificial Neural Networks," in *2009 International Association of Computer Science and Information Technology - Spring Conference*, 2009, pp. 456–458.

[41] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Comput. Methods Programs Biomed.*, vol. 108, no. 2, pp. 570–579, 2012.

[42] I. Hamamoto, S. Okada, T. Hashimoto, H. Wakabayashi, T. Maeba, and H. Maeta, "Prediction of the early prognosis of the hepatectomized patient with hepatocellular carcinoma with a neural network.," *Comput. Biol. Med.*, vol. 25, no. 1, pp. 49–59, Jan. 1995.

[43] K. Mala and V. Sadasivam, "Automatic Segmentation and Classification of Diffused Liver Diseases using Wavelet Based Texture Analysis and Neural Network," in *2005 Annual IEEE India Conference - Indicon*, pp. 216–219.

[44] K. Revett, F. Gorunescu, M. Gorunescu, and M. Ene, "Mining A Primary Biliary Cirrhosis Dataset Using Rough Sets and a Probabilistic Neural Network," in *2006 3rd International IEEE Conference Intelligent Systems*, 2006, pp. 284–289.

[45] Y. Dong, Z. Xia, and Z. Xia, "A two-level approach to choose the cost parameter in support vector machines," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 1366–1370, 2008.

[46] C. T. Su and C. H. Yang, "Feature selection for the SVM: An application to hypertension diagnosis," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 754–763, 2008.

[47] G. S. Uttreshwar and A. A. Ghatol, "Hepatitis B Diagnosis Using Logical Inference And Generalized Regression Neural Networks," in *2009 IEEE International Advance Computing Conference*, 2009, pp. 1587–1595.

[48] C. T. C. Arsene and P. J. Lisboa, “Bayesian Neural Network Applied in Medical Survival Analysis of Primary Biliary Cirrhosis,” pp. 81–85, 2012.

[49] G. S. Babu and S. Suresh, “Meta-cognitive RBF Network and its Projection Based Learning algorithm for classification problems,” *Appl. Soft Comput.*, vol. 13, no. 1, pp. 654–666, 2013.

[50] J. H. Jeon, J. Y. Choi, S. Lee, and Y. M. Ro, “Multiple ROI selection based focal liver lesion classification in ultrasound images,” *Expert Syst. Appl.*, vol. 40, no. 2, pp. 450–457, 2013.

[51] A. G. Floares, “Intelligent clinical decision supports for interferon treatment in chronic hepatitis C and B based on i-biopsy&#x2122,” in *2009 International Joint Conference on Neural Networks*, 2009, pp. 855–860.

[52] W. Yan, M. Lizhuang, L. Xiaowei, and L. Ping, “Correlation between Child-Pugh Degree and the Four Examinations of Traditional Chinese Medicine (TCM) with Liver Cirrhosis,” *2008 Int. Conf. Biomed. Eng. Informatics*, pp. 858–862, 2008.

[53] M. Eastwood and B. Gabrys, “Generalised bottom-up pruning: A model level combination of decision trees,” *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9150–9158, 2012.

[54] S. Kohara, T. Tateyama, A. H. Foruzan, A. Furukawa, S. Kanasaki, M. Wakamiya, and Y. Chen, “Application of statistical shape model to diagnosis of liver disease,” *Softw. Eng. Data Min. (SEDM), 2010 2nd Int. Conf.*, no. 1, pp. 680–683, 2010.

[55] Y. Luo, J. Q. Li, D. W. Zheng, Z. P. Tan, H. Zhou, Q. P. Deng, Y. T. Liu, A. H. Ou, and J. Yin, “Application of data mining technology in excavating prevention and treatment experience of infectious diseases from famous herbalist doctors,” in *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2011*, 2011, pp. 784–790.

[56] C. H. Jen, C. C. Wang, B. C. Jiang, Y. H. Chu, and M. S. Chen, “Application of classification techniques on development an early-warning system for chronic illnesses,” *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8852–8858, 2012.

[57] M. Neshat, M. Yaghobi, M. B. Naghibi, and A. Esmaelzadeh, “Fuzzy expert system design for diagnosis of liver disorders,” in *Proceedings - 2008 International Symposium on Knowledge Acquisition and Modeling, KAM 2008*, 2008, pp. 252–256.

[58] I. Gadaras and L. Mikhailov, “An interpretable fuzzy rule-based classification methodology for medical diagnosis,” *Artif. Intell. Med.*, vol. 47, no. 1, pp. 25–41, 2009.

[59] O. U. Obot and S. S. Udoh, “A framework for fuzzy diagnosis of hepatitis,” *2011*

World Congr. Inf. Commun. Technol., pp. 439–443, 2011.

[60] L. K. Ming, L. C. Kiong, and L. W. Soong, "Autonomous and deterministic supervised fuzzy clustering with data imputation capabilities," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 1117–1125, 2011.

[61] A. M. Badawi, A. S. Derbala, and A. B. M. Youssef, "Fuzzy logic algorithm for quantitative tissue characterization of diffuse liver diseases from ultrasound images," *Int. J. Med. Inform.*, vol. 55, no. 2, pp. 135–147, 1999.

[62] P. Luukka, "Fuzzy beans in classification," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 4798–4801, 2011.

[63] S. Ozsen and S. Gunes, "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 386–392, 2009.

[64] Y. Zhang and P. I. Rockett, "A Generic Optimising Feature Extraction Method Using Multiobjective Genetic Programming," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 1087–1097, 2011.

[65] K. Tan, Q. Yu, C. Heng, and T. Lee, "Evolutionary Computing for Knowledge Discovery in Medical Diagnosis," *Artif. Intell. Med.*, 2003.

[66] I. De Falco, "Differential Evolution for automatic rule extraction from medical databases," *Appl. Soft Comput. J.*, vol. 13, no. 2, pp. 1265–1283, 2013.

[67] C.-C. Wu, W.-L. Lee, Y.-C. Chen, C.-H. Lai, and K.-S. Hsieh, "Ultrasonic liver tissue characterization by feature fusion," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9389–9397, 2012.

[68] G. Bologna, "A model for single and multiple knowledge based networks," *Artif. Intell. Med.*, vol. 28, no. 2, pp. 141–163, 2003.

[69] D. Calisir and E. Dogantekin, "A new intelligent hepatitis diagnosis system: PCA–LSSVM," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10705–10708, Aug. 2011.

[70] A. M. Hashem, M. E. M. Rasmy, K. M. Wahba, and O. G. Shaker, "Single stage and multistage classification models for the prediction of liver fibrosis degree in patients with chronic hepatitis C infection," *Comput. Methods Programs Biomed.*, vol. 105, no. 3, pp. 194–209, 2012.

[71] E. Comak, K. Polat, S. Gunes, and A. Arslan, "A new medical decision making system: Least square support vector machine (LSSVM) with Fuzzy Weighting Pre-processing," *Expert Syst. Appl.*, vol. 32, no. 2, pp. 409–414, 2007.

[72] M. Neshat and A. E. Zadeh, "Hopfield neural network and fuzzy Hopfield neural

network for diagnosis of liver disorders," in *2010 5th IEEE International Conference Intelligent Systems*, 2010, pp. 162–167.

[73] E. Dogantekin, A. Dogantekin, and D. Avci, "Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System," *Expert Syst. Appl.*, vol. 36, no. 8, pp. 11282–11286, 2009.

[74] A. Celikyilmaz, I. B. Turksen, R. Aktas, M. M. Doganay, and N. B. Ceylan, "Increasing accuracy of two-class pattern recognition with enhanced fuzzy functions," *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 1337–1354, 2009.

[75] R. Ceylan, M. Ceylan, Y. Ozbay, and S. Kara, "Fuzzy clustering complex-valued neural network to diagnose cirrhosis disease," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9744–9751, 2011.

[76] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets.," *Comput. Biol. Med.*, vol. 40, no. 5, pp. 509–518, 2010.

[77] D.-C. Li and C.-W. Liu, "A class possibility based kernel to increase classification accuracy for small data sets using support vector machines," *Expert Syst. Appl.*, vol. 37, no. 4, pp. 3104–3110, Apr. 2010.

[78] S. Kulluk, L. Ozbakır, and A. Baykasoglu, "Fuzzy DIFACONN-miner: A novel approach for fuzzy rule extraction from neural networks," *Expert Syst. Appl.*, vol. 40, no. 3, pp. 938–946, 2013.

[79] K. Polat, S. Sahan, H. Kodaz, and S. Gunes, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism," *Expert Syst. Appl.*, vol. 32, no. 1, pp. 172–183, 2007.

[80] E. Mezyk and O. Unold, "Mining fuzzy rules using an Artificial Immune System with fuzzy partition learning," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 1965–1974, 2011.

[81] F. Gorunescu, S. Belciug, M. Gorunescu, and R. Badea, "Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 12824–12832, 2012.

[82] S. Dehuri and S. B. Cho, "Evolutionarily optimized features in functional link neural network for classification," *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4379–4391, 2010.

[83] S. N. Qasem and S. M. Shamsuddin, "Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 1427–1438, 2011.

[84] R.-H. Lin, "An intelligent model for liver disease diagnosis," *Artif. Intell. Med.*, vol.

47, no. 1, pp. 53–62, 2009.

[85] Y. J. Park, S. H. Chun, and B. C. Kim, "Cost-sensitive case-based reasoning using a genetic algorithm: application to medical diagnosis," *Artif Intell Med*, vol. 51, pp. 133–45, 2011.

[86] B. K. Sarkar, S. S. Sana, and K. Chaudhuri, "A genetic algorithm-based rule extraction system," *Appl. Soft Comput.*, vol. 12, no. 1, pp. 238–254, 2012.

[87] C. Stoean, R. Stoean, M. Lupsor, H. Stefanescu, and R. Badea, "Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis," *Comput. Biol. Med.*, vol. 41, no. 4, pp. 238–246, 2011.

[88] P. Luukka and T. Leppalampi, "Similarity classifier with generalized mean applied to medical data," *Comput. Biol. Med.*, vol. 36, no. 9, pp. 1026–1040, 2006.

[89] P. Luukka, "Classification based on fuzzy robust PCA algorithms and similarity classifier," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7463–7468, 2009.

[90] Y. Torun and G. Tohumoglu, "Designing simulated annealing and subtractive clustering based fuzzy classifier," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2193–2201, 2011.

[91] C. H. Wang, T. P. Hong, and S. S. Tseng, "Integrating fuzzy knowledge by genetic algorithms," *IEEE Trans. Evol. Comput.*, vol. 2, no. 4, pp. 138–149, 1998.

[92] N. A. Chowdhury, M. Khatun, and M. M. A. Hashem, "On integrating fuzzy knowledge using a novel evolutionary algorithm," in *2007 10th International Conference on Computer and Information Technology, ICCIT*, 2007.

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

[94] O. U. Obot and F. M. E. Uzoka, "A framework for application of neuro-case-rule base hybridization in medical diagnosis," *Appl. Soft Comput. J.*, vol. 9, no. 1, pp. 245–253, 2009.

[95] K. Polat and S. Gunes, "A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS.," *Comput. Methods Programs Biomed.*, vol. 88, no. 2, pp. 164–74, 2007.

[96] K. Polat and S. Gunes, "Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection," *Expert Syst. Appl.*, vol. 33, no. 2, pp. 484–490, 2007.

[97] C. T. Su, L. S. Chen, and Y. Yih, "Knowledge acquisition through information granulation for imbalanced data," *Expert Syst. Appl.*, vol. 31, no. 3, pp. 531–541, 2006.

[98] D.-C. Li, C.-W. Liu, and S. C. Hu, “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.

[99] R. Stoean, C. Stoean, M. Lupsor, H. Stefanescu, and R. Badea, “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, 2015.

[100] M. Cruz-Ramírez, C. Hervas-Martínez, J. C. Fernández, J. Briceño, and M. De La Mata, “Multi-objective evolutionary algorithm for donor-recipient decision system in liver transplants,” *Eur. J. Oper. Res.*, vol. 222, no. 2, pp. 317–327, 2012.

[101] P. C. Chang, J. J. Lin, and C. H. Liu, “An attribute weight assignment and particle swarm optimization algorithm for medical database classifications,” *Comput. Methods Programs Biomed.*, vol. 107, no. 3, pp. 382–392, 2012.

[102] K.-S. Leung, K. H. Lee, J.-F. Wang, E. Y. T. Ng, H. L. Y. Chan, S. K. W. Tsui, T. S. K. Mok, P. C.-H. Tse, and J. J.-Y. Sung, “Data mining on DNA sequences of hepatitis B virus.,” *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 2, pp. 428–40, 2011.

[103] P. C. Chang, C. Y. Fan, and W. Y. Dzan, “A CBR-based fuzzy decision tree approach for database classification,” *Expert Syst. Appl.*, vol. 37, no. 1, pp. 214–225, 2010.

[104] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. C. Hsieh, “A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification,” *Appl. Soft Comput.*, vol. 11, no. 1, pp. 632–644, 2011.

[105] Y. Guo, T. Hastie, and R. Tibshirani, “Regularized linear discriminant analysis and its application in microarrays.,” *Biostatistics*, vol. 8, no. 1, pp. 86–100, Jan. 2007.

[106] J. Ye and Q. Li, “A two-stage linear discriminant analysis via QR-decomposition.,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 929–41, Jun. 2005.

[107] S. Noushath, G. Hemantha Kumar, and P. Shivakumara, “Diagonal Fisher linear discriminant analysis for efficient face recognition,” *Neurocomputing*, vol. 69, no. 13–15, pp. 1711–1716, 2006.

[108] S. Srivastava, M. R. Gupta, and B. A. Frigyik, “Bayesian Quadratic Discriminant Analysis,” *J. Mach. Learn. Res.*, vol. 8, pp. 1277–1305, Dec. 2007.

[109] J. R. Schott, “Dimensionality reduction in quadratic discriminant analysis,” *Comput. Stat. Data Anal.*, vol. 16, no. 2, pp. 161–174, 1993.

[110] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[111] J. A. K. Suykens and J. Vandewalle, “Least Squares Support Vector Machine

Classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.

[112] D. Tsujinishi and S. Abe, "Fuzzy least squares support vector machines for multiclass problems," in *Neural Networks*, 2003, vol. 16, no. 5–6, pp. 785–792.

[113] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010*, 2010, vol. 1, pp. 91–94.

[114] H. Samet, "K-nearest neighbor finding using MaxNearestDist," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 243–252, 2008.

[115] W. Dong, C. Moses, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in *WWW*, 2011, pp. 577–586.

[116] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[117] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Syst. Appl.*, vol. 67, 2017.

[118] E. Elyan and M. M. Gaber, "A genetic algorithm approach to optimising random forests applied to class engineered data," *Inf. Sci. (Ny).*, vol. 384, pp. 220–234, 2017.

[119] F. Pourpanah, C. J. Tan, C. P. Lim, and J. Mohamad-Saleh, "A Q-learning-based multi-agent system for data classification," *Appl. Soft Comput. J.*, vol. 52, pp. 519–531, 2017.

[120] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Inform.*, vol. 59, pp. 185–200, 2016.

[121] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "HMV: A medical decision support framework using multi-layer classifiers for disease prediction," *J. Comput. Sci.*, vol. 13, pp. 10–25, 2016.

[122] H. K. Sok, M. P. L. Ooi, Y. C. Kuang, and S. Demidenko, "Multivariate alternating decision trees," *Pattern Recognit.*, vol. 50, pp. 195–209, 2016.

[123] P. C. Zou, J. Wang, S. Chen, and H. Chen, "Margin distribution explanation on metric learning for nearest neighbor classification," *Neurocomputing*, vol. 177, pp. 168–178, 2016.

[124] M. ABDAR, "A Survey and Compare the Performance of IBM SPSS Modeler and Rapid Miner Software for Predicting Liver Disease by Using Various Data Mining Algorithms," *Cumhur. Sci. J.*, vol. 36, no. 3, pp. 3230–3241, 2015.

[125] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, and T. Ing Ren, "META-DES: A dynamic ensemble selection framework using meta-learning," *Pattern Recognit.*, vol. 48, no. 5, pp. 1925–1935, 2015.

[126] H. K. Sok, M. P. L. Ooi, and Y. C. Kuang, "Sparse alternating decision tree," *Pattern Recognit. Lett.*, vol. 60–61, pp. 57–64, 2015.

[127] L. V. Utkin and A. I. Chekh, "A new robust model of one-class classification by interval-valued training data using the triangular kernel," *Neural Networks*, vol. 69, pp. 99–110, 2015.

[128] Y. Xu, J. Yu, and Y. Zhang, "KNN-based weighted rough ν-twin support vector machine," *Knowledge-Based Syst.*, vol. 71, pp. 303–313, 2014.

[129] A. Yan, L. Qian, and C. Zhang, "Memory and forgetting: An improved dynamic maintenance method for case-based reasoning," *Inf. Sci. (Ny).*, vol. 287, pp. 50–60, 2014.

[130] H. Jin, S. Kim, and J. Kim, "Decision Factors on Effective Liver Patient Data Prediction," *Int. J. Bio-Science Bio-Technology*, vol. 6, no. 4, pp. 167–178, 2014.

[131] B. Venkata Ramana, M. S. P. Babu, and N. . Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *Int. J. Database Manag. Syst.*, vol. 3, no. 2, pp. 101–114, 2011.

[132] E. J. Carey, A. H. Ali, and K. D. Lindor, "Primary biliary cirrhosis," *Lancet*, vol. 386, no. 10003, pp. 1565–1575, 2015.

[133] J. K. Dyson, G. M. Hirschfield, D. H. Adams, U. Beuers, D. A. Mann, K. D. Lindor, and D. E. J. Jones, "Novel therapeutic targets in primary biliary cirrhosis.," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 12, no. 3, pp. 147–58, 2015.

[134] R. Poupon, "Primary biliary cirrhosis: A 2010 update," *Journal of Hepatology*, vol. 52, no. 5. pp. 745–758, 2010.

[135] T. Therneau and P. Grambsch, "Modeling Survival Data: Extending the Cox Model," *Technometrics*, vol. 44, no. 1, pp. 85–86, 2002.

[136] R. Banerjee, M. Pavlides, E. M. Tunnicliffe, S. K. Piechnik, N. Sarania, R. Philips, J. D. Collier, J. C. Booth, J. E. Schneider, L. M. Wang, D. W. Delaney, K. A. Fleming, M. D. Robson, E. Barnes, and S. Neubauer, "Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease," *J. Hepatol.*, vol. 60, no. 1, pp. 69–77, 2014.

[137] T. Van Erven and P. Harrëmos, "Rényi divergence and kullback-leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[138] J. Shlens, “A Tutorial on Principal Component Analysis,” *ArXiv*, pp. 1–13, 2014.

[139] M. Ringnér, “What is principal component analysis?,” *Nat. Biotechnol.*, vol. 26, no. 3, pp. 303–304, 2008.

[140] R. Bro and A. K. Smilde, “Principal component analysis,” *Anal. Methods*, vol. 6, no. 9, p. 2812, 2014.

[141] D. T. Larose, “k-Nearest Neighbor Algorithm,” in *Discovering Knowledge in Data: An Introduction to Data Mining*, 2004, pp. 90–106.

[142] S. Balakrishnama and a Ganapathiraju, “Linear Discriminant Analysis - a Brief Tutorial,” *Compute*, vol. 11, pp. 1–9, 1998.

[143] M. Welling, “Fisher Linear Discriminant Analysis,” *Science (80-.).*, vol. 1, no. 2, pp. 1–3, 2009.

[144] J. Hua, Z. Xiong, and E. R. Dougherty, “Determination of the optimal number of features for quadratic discriminant analysis via the normal approximation to the discriminant distribution,” *Pattern Recognit.*, vol. 38, no. 3, pp. 403–421, 2005.

[145] A. Moore, “K-means and Hierarchical Clustering,” *Stat. Data Min. Tutorials*, pp. 1–24, 2001.

[146] H. Xiong, J. Wu, and J. Chen, “K-means clustering versus validation measures: A data-distribution perspective,” *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 39, no. 2, pp. 318–331, 2009.

[147] D. Steinley, “K-means clustering: a half-century synthesis.,” *Br. J. Math. Stat. Psychol.*, vol. 59, no. Pt 1, pp. 1–34, 2006.

[148] R. E. Schapire, “Explaining adaboost,” in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 2013, pp. 37–52.

[149] R. E. Schapire, “The boosting approach to machine learning: an overview,” *Nonlinear Estim. Classif.*, vol. 171, pp. 149–171, 2003.

[150] R. Rojas, “AdaBoost and the Super Bowl of Classifiers A Tutorial Introduction to Adaptive Boosting,” *Writing*, pp. 1–6, 2009.

[151] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, 1999.

[152] R. Maclin and D. Opitz, “An Empirical Evaluation of Bagging and Boosting,” *Proc. Fourteenth Natl. Conf. Artif. Intell.*, no. September, pp. 546–551, 1997.

[153] J. Quinlan, “Bagging, boosting, and C4. 5,” *Proc. Thirteen. Natl. Conf. Artif. Intell.*, pp. 725–730, 1996.

[154] S. B. Kotsiantis and P. E. Pintelas, “Combining Bagging and Boosting,” *Comput.*

Intell., vol. 1, no. 4, pp. 324–333, 2004.

[155] E. Bauer, R. Kohavi, P. Chan, S. Stolfo, and D. Wolpert, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Mach. Learn.*, vol. 36, no. August, pp. 105–139, 1999.

[156] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *Int. J. Adv. Comput. Sci. Appl.*, no. 2, pp. 13–19, 2014.

[157] K. Polat and S. Gunes, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems," *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 1587–1592, 2009.

[158] S. Drazin, "Decision Tree Analysis using Weka," *Mach. Learn. II, Univ. Miami*, pp. 1–3, 2010.

[159] J. R. Quinlan, "Bagging, boosting, and C4.5," *Proc. Thirteen. Natl. Conf. Artif. Intell.*, vol. 5, no. Quinlan 1993, pp. 725–730, 2006.

[160] W. Dich, K. Grudzinski, and G. H. F. Diercksen, "Minimal distance neural methods," *1998 IEEE Int. Jt. Conf. Neural Networks Proceedings. IEEE World Congr. Comput. Intell. (Cat. No.98CH36227)*, vol. 2, 1998.

[161] W. Duch, R. Adamczak, and K. Grabczewski, "Optimization of logical rules derived by neural procedures," *IJCNN'99. Int. Jt. Conf. Neural Networks. Proc. (Cat. No.99CH36339)*, vol. 1, 1999.

[162] B. Ster and A. Dobnikar, "Neural networks in medical diagnosis: comparison with other methods," *Solving Eng. Probl. with Neural Networks. Proc. Int. Conf. Eng. Appl. Neural Networks (EANN'96). Syst. Eng. Assoc, Turku, Finl.*, vol. 1, no. April 2016, pp. 427–430, 1996.

[163] K. Polat and S. Gunes, "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation," *Digit. Signal Process.*, vol. 16, no. 6, pp. 889–901, 2006.

[164] M. S. Bascil and F. Temurtas, "A study on hepatitis disease diagnosis using multilayer neural network with Levenberg Marquardt training algorithm," *J. Med. Syst.*, vol. 35, no. 3, pp. 433–436, 2011.

[165] K. C. Tan, E. J. Teoh, Q. Yu, and K. C. Goh, "A hybrid evolutionary algorithm for attribute selection in data mining," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8616–8630, 2009.

[166] M. H. Zangooei, J. Habibi, and R. Alizadehsani, "Disease Diagnosis with a hybrid method SVR using NSGA-II," *Neurocomputing*, vol. 136, pp. 14–29, 2014.

[167] B. Naik, J. Nayak, H. S. Behera, and A. Abraham, "A self adaptive harmony search based functional link higher order ANN for non-linear data classification," *Neurocomputing*, vol. 179, 2016.

[168] S. Kulluk, L. Ozbakir, P. Z. Tapkan, and A. Baykasoğlu, "Cost-sensitive meta-learning classifiers: MEPAR-miner and DIFACONN-miner," *Knowledge-Based Systems*, 2015.

[169] Q. Hou, L. Zhen, N. Deng, and L. Jing, "Novel Grouping Method-based support vector machine plus for structured data," *Neurocomputing*, vol. 211, pp. 191–201, 2016.

[170] Y. Hayashi and K. Fukunaga, "Accuracy of rule extraction using a recursive-rule extraction algorithm with continuous attributes combined with a sampling selection technique for the diagnosis of liver disease," *Informatics Med. Unlocked*, vol. 5, pp. 26–38, 2016.

[171] T. Kanik, "Hepatitis disease diagnosis using Rough Set - modification of the pre-processing algorithm," in *Information and Communication Technologies - International Conference*, 2012, pp. 47–50.

[172] R. S. O'Shea, S. Dasarathy, and A. J. McCullough, "Alcoholic liver disease.," *Hepatology*, vol. 51, no. 1, pp. 307–328, 2010.

[173] A. Louvet and P. Mathurin, "Alcoholic liver disease: mechanisms of injury and targeted treatment.," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 12, no. 4, pp. 231–42, 2015.

[174] W. C. Zhou, Q. B. Zhang, and L. Qiao, "Pathogenesis of liver cirrhosis," *World J. Gastroenterol.*, vol. 20, no. 23, pp. 7312–7324, 2014.

[175] E. A. Tsochatzis, J. Bosch, and A. K. Burroughs, "Liver cirrhosis.," *Lancet (London, England)*, vol. 383, no. 9930, pp. 1749–61, 2014.

[176] C. E. Broelsch, A. Frilling, and M. Malago, "Hepatoma - Resection or transplantation," *Surgical Clinics of North America*, vol. 84, no. 2. pp. 495–511, 2004.

[177] T. Y. Luo and N. Q. Pan, "A study on the multiple factorial causation of hepatoma among young patients," *Zhonghua Liu Xing Bing Xue Za Zhi*, vol. 18, no. 1, pp. 37–39, 1997.

[178] J. L. Thistle, P. A. Cleary, J. M. Lachin, M. P. Tyor, and T. Hersh, "The natural history of cholelithiasis: The National Cooperative Gallstone Study," *Ann. Intern. Med.*, vol. 101, no. 2, pp. 171–175, 1984.

[179] D. E. Cafasso and R. R. Smith, "Symptomatic cholelithiasis and functional disorders of the biliary tract," *Surgical Clinics of North America*, vol. 94, no. 2. pp. 233–256, 2014.

[180] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Hierarchical Clustering*. 2011.

[181] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012.

[182] Y. Qi, “Random forest for bioinformatics,” in *Ensemble Machine Learning: Methods and ApplicatiOns*, 2012, pp. 307–323.

[183] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. K??nig, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 493–507, 2012.

[184] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[185] S. K. Pal and S. Mitra, “Multilayer Perceptron, Fuzzy Sets, and Classification,” *IEEE Trans. Neural Networks*, vol. 3, no. 5, pp. 683–697, 1992.

[186] A. Niyom, S. Chiewchanwattana, K. Sunat, and C. Lursinsap, “The DAGs-MLP Structure to the Efficiency of Neural Network Classification for Diagnosis of Hepatobiliary Disoders,” in *2006 SICE-ICASE International Joint Conference*, 2006, pp. 1489–1493.

[187] K. Ooba and E. Tazaki, “Leaning of neural networks using extended genetic algorithm with neutral mutation-application to ill-posed medical diagnostic system [for leaning read learning],” in *IJCNN’99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, vol. 6, pp. 4161–4165.

[188] Y. Hayashi, R. Setiono, and K. Yoshida, “Diagnosis of hepatobiliary disorders using rules extracted from artificial neural networks,” in *IEEE International Conference on Fuzzy Systems*, 1999, vol. 1.

[189] T. Ichimura, E. Tazaki, and K. Yoshida, “Reasoning and learning method for fuzzy rules using neural networks and its application to medical diagnosis,” in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*, vol. 6, pp. 3534–3539.

[190] T. Ichimura, S. Oeda, and K. Yoshida, “An adaptive evolutional neuro learning method using genetic search and extraction of rules from trained networks,” in *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*, vol. 2, pp. 1343–1350.