Republic of Iraq

Ministry of Higher Education and Scientific Research

University of Diyala

Diagnosis Of Liver Disease Using Machine Learning Algorithms

A Thesis
Submitted to the Department of Computer Science\ College of Science\ University of Diyala in a Partial Fulfilment of the Requirements for the Degree of Master in Computer Science

By

Sondos Jameel Mukhyber

Supervised By

Prof. Dr. Dhahir Abdulhade Abdulah
Prof. Dr. Amer Dawood Majeed

2021 A.D. 1442 A.H.

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

﴿يَرْفَعِ اللَّهُ الَّذِينَ آمَنُوا مِنكُمْ
وَالَّذِينَ أُوتُوا الْعِلْمَ دَرَجَاتٍ﴾

صدق الله العظيم

(سورة المجادلة الاية:11)

Acknowledgment

*Firstly, all my prayers go to **Allah**, the Almighty, who helped me to accomplish this work until its end.*

*I wish to express my thanks to my college (college of science), my supervisors, **Prof. Dr. Dhahir Abdulhade Abdulah and Prof. Dr. Amer Dawood Majeed** for supervising this research, and for the generosity, patience, and continuous guidance throughout the work. It has been my good fortune to have advice and guidance from them. My thanks to the academic and administrative staff at the Department of the computer sciences.*

I would like to thank my family who has supported me throughout and encouraging me to further my studies and help to complete this project, and I wish to express my thanks to my friends.

Sondos jameel

Dedication

I would like to dedicate this work to:

Whom taught me that the champions will never be defeated, but they convert it to victory.

Our Prophet Mohammed

Peace be Upon Him (PBH)

My Father, My Mother, My Sisters, and My Brothers.

Sondos jameel

Abstract

In the present time, liver disease becomes one of the significant topics for researchers in the field of medical technology. Patients with liver disease can only be saved if and only if the disease is early discovered, otherwise the chance of survival chances is slim. The most significant issue, according to this, is early and accurate detection, which is crucial for curing this disease. A Classification is one of the main issues in the knowledge discovery fields and sciences decision. There are many types of algorithms used for constructing classifiers and liver disorder detection.

In this thesis, efficient liver disease classification model have been built to increase the accuracy and decrease the error rate in the disorder detection process. The proposed model consists of four main stages: pre-processing, split of data, z- score normalization and classification of the liver disease stage. Datasets may contain some missing values. In this proposed model the way to handle these values is done by replacing each missing value with the mean value. In this thesis, five algorithm (Decision Tree, Random Forest, Artificial Neural Networks, Support Vector Machine, and K-Nearest Neighbour) are used.

The proposed model has been tested by using two liver datasets (Indian Liver patients and Iraqi liver patients). In comparison to other existing approaches, the results show that the model has a high accuracy rate, the accuracy of the Decision Tree, Random Forest, Artificial Neural Networks, Support Vector Machine, and K-Nearest Neighbour using Iraqi liver patient dataset is 99.06%, 99.06%, 95.32%, 89.71, and 87.85% and when using Indian Liver patient Dataset the accuracy rate is 80.34%, 77.77%, 84.61%, 80.34%, and 77.77%.

List of Contents

List of Abbreviations

Abbreviations	Description
ANN	Artificial Neural Network
AST	Aspartate aminotransferase
ALT	Alanine transaminase
ALP	Alkaline phosphatase
CLS	Conceptual Learning System
DT	Decision Tree
ID3	Iterative DiChaudomiser3
ILPD	Indian Liver Patient Dataset
K-NN	K-Nearest Neighbor
KDD	Knowledge Discovery Databases
RF	Random Forest
SVM	Support Vector Machine

List of Figures

List of Tables

List of Algorithms

Chapter One
General Introduction

Chapter one

General Introduction

1.1 Introduction

The Healthcare industry contains big and complex data that may be required in order to discover fascinating pattern of diseases and makes effective decisions with the help of different machine learning techniques. data mining techniques are used to discover knowledge in database and for medical research [1].

Data mining techniques such as classification and prediction, clustering, association rule mining and various mining methods can be useful to apply on medical data.

liver has a vital role in the human body functions from protein production to removing toxins from the body and it is essential for the survival. The failure of liver functioning leads to the death. The functioning of the liver is examined by two types of tests such as imaging test and liver function tests which help to diagnose liver diseases. Liver diseases are caused by many factors such as stress, food habits, consumption of alcohol drug intake, etc. In recent days, it could be found that it is very difficult to detect at an early stages symptoms are very hard to identify. The physician often slips to detect the liver disease which leads to improper medical treatment. Various data mining algorithms can be used to classify the various disease stages including early stage so that it could be help the physician to give the proper treatment [2].

Classification can be considered as one of the most common operations in data mining. Classification is a process that divides the dataset into specified sections and then classifies the data which is a two-phase process:

In the first phase, it develops a model based on educational datasets of databases and then creates educational dataset including records, samples, examples and things with a collection of attributes and aspects. Each sample has a specific class label [3].

In the second phase, the developed model in the previous phase is used to classify new samples [3].

Classification algorithms can be either supervised or unsupervised based on the learning mechanism. Supervised learning is implemented by set of labels defined prior in the training set. The function is mapped for new unseen data to predict the labels. Few examples are Artificial Neural Network, Bagging, Boosting, Naive Bayes, Kernel-based classifiers, Nearest Neighbor algorithm, Decision Trees, Random Forest, and other ensemble of classifiers. Whereas unsupervised learning identifies the missing or hidden patterns in unlabelled data without any labels. They are commonly used for dimensionality reduction of feature space. The unsupervised ensembles include clustering approaches, self-organization maps, hidden Markov models and adaptive resonance theory [4].

In this thesis, a liver disease classification system is proposed, which uses the liver function test of a possible patient along with his/her private information. Each attribute of liver function test is normalized its values using z-score normalization. The Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) algorithms is used as a classifier to classify if the person have liver disease or not

1.2 Related Works

This section reviews some of previous studies and explains the different techniques that are used for classify liver disease.

- **L. Alice Auxilia (2018) [5]:** proposed a paper based on different classification techniques such as Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine and Artificial Neural Network. The dataset are taken from the UCI vault. It has been seen that decision tree gives better outcomes contrast with other order classification algorithms. The accuracy obtained of applying the Decision Tree Random Forest, Support Vector Machine and Artificial Neural Network was 81%, 77%, 77%, and 71%.
- **Nazmun Nahar & Ferdous Ara (2018) [6]:** proposed a paper for prediction liver disease in early stages by using different decision tree techniques J48, LMT, Random Forest, Random tree, REPTree, Decision Stump & Hoeffding Tree. The data are collected from UCI Machine Learning Repository. The analysis proves that Decision Stump provides the highest accuracy than other

techniques. Decision Stump outperforms well than other algorithms and its achieved accuracy is 70.67% while the accuracy with Random Forest is 69.30%.

- **Thirunavukkarasu K et al. (2018) [7]:** proposed a paper to predict liver disease using different classification algorithms. The algorithms used for this purpose of work is Logistic Regression, K-Nearest Neighbour and Support Vector Machines. The dataset is taken is from the Indian Liver Patient Dataset (ILPD). This is downloaded from UCL Machine Learning Repository. Accuracy score and confusion matrix is used to compare this classification algorithm. Logistic Regression and K-Nearest Neighbour have the highest accuracy with 73.97% while the accuracy with Support Vector Machines is 71.97%.
- **Nazim Razali et al. (2019) [8]:** proposed paper to predict liver disease using different classification algorithms. This study used classification and regression as data mining tasks. In classification, models are evaluated using Bayes point machine and neural network as algorithms, while regression using linear regression and Poisson regression. This study used dataset from the University of California Irvine (UCI) repository. The results showed that Bayes Point Machines algorithm is the best algorithm used for solve the problem relating to the liver disease with accuracy 70% while neural network achieved accuracy with 66%.
- **A.K.M Sazzadur Rahman et al. (2019) [9]:** proposed paper to evaluate the performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. They collected a dataset from the UCI Machine Learning Repository. They used six algorithms Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated and The accuracy that obtained was 75%, 74%, 69%, 64%, 62% and 53% for LR, RF, DT, SVM, KNN and NB.
- **Md. Shafiul Azam et al. (2020) [10]:** proposed paper to predict liver disease using some efficient classification algorithms: Random Forest, Perceptron, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). These classification

algorithms are applied to the original liver patient datasets collected from the UCI repository, then analysed features and tweaked to improve the performance of this predictor and made a comparative analysis among the classifiers. The accuracy before Apply feature selection was 60%, 39%, 64%, 66%, and 71% for DT, P, RF, KNN, and SVM. The accuracy after apply feature selection became 72%, 66%, 73%, 74%, and 72%.

- **Rashid Naseem et al. (2020) [11]**: In this work, Researchers are trying to project a model for the early prediction of liver disease utilizing various machine learning approaches. They used ten classifiers including A1DE, NB, MLP, SVM, KNN, CHIRP, CDT, Forest-PA, J48, and RF to find the optimal solution for early and accurate prediction of liver disease. The datasets utilized in this study are taken from the UCI ML repository and the GitHub (ILPD) repository. It has been seen that the RF classifier has highest accuracy for dataset from UCI ML repository and accuracy was 72.17%,58.26% and 62,89% for RF, SVM, KNN. SVM has highest accuracy for Dataset from GitHub repository (ILPD) and accuracy was 71.35%, 69.29%, and 64.15% for SVM, RF, and KNN.
- **Jagdeep Singh et al. (2020) [12]**: proposed paper to predict liver disease using classification and feature selection technique. The implementation of proposed work is done using (ILPD) using different classification algorithms such as such as Random Forest, K-Nearest Neighbor, SMO, Logistic Regression, Naïve Bayes, and J48. The accuracy was achieved by RF, and KNN before using feature selection technique 71.53%, and 64.15%, while the accuracy after apply feature selection technique 71.87%, and 67.41%.

1.3 Problem Statement

Patients with the liver disease continue to increase and the symptoms of the disease are difficult to detect. Therefore many people suffer from liver damage but they feel healthy, it causes many medical practitioners to often fail to detect the disease. Therefore accurate detection is necessary to help the medical practitioner to give proper medication and medical treatment. So the medical classification model will help a doctor to automatic classification and reduces the workload.

1.4 Aim of Thesis

This thesis aims to design and implement a liver disease classification model able to accurately classify if the person have liver disease or not based on the 10 important attributes of liver disease using a DT, RF, ANN, SVM, and KNN.

1.5 Outline of Thesis

The rest chapters in this thesis are organized as follows:

Chapter Two: Theoretical Background

This chapter clarifies the definition and Knowledge discovery in databases (KDD) steps, it also explains the Data mining techniques used for classification.

Chapter Three: The Proposed System

This chapter describes the proposed classification model with its design and implementation.

Chapter Four: Experimental Results and Evaluation

This chapter shows the implementation results of the proposed model steps and evaluates these results.

Chapter Five: Conclusions and Suggestions for Future work

This chapter presents the conclusions of this work. Furthermore, it provides suggestions for future work.